Gagnasafnsfræði

Páll Melsted

25. sept.

Hönnun gagnagrunna

Tengsl milli gagna ákvarða hvernig er "best" að hanna gagnagrunna.

1. Hvernig á að skipuleggja gagnagrunn upp í töflur
2. Hvernig er hægt að taka lélegan gagnagrunn og laga
3. Formlegar aðferðir við að brjóta upp töflur.

Fallákveður

Aðaltólið sem við notum til þess að tala um tengls milli gagna eru fallákveður (functional dependencies, FD).

Eiginleikar $\bar{A} = A_1, \ldots, A_n$ ákvarða $\bar{B} = B_1, \ldots, B_m$ í R ef allar n-dir með sömu gildi fyrir $A_1, \ldots, A_n$ hafa sömu gildi fyrir $B_1, \ldots, B_m$. Við táknum þetta sem

$$A_1, \ldots, A_n \to B_1, \ldots, B_m$$

eða

$$\bar{A} \to \bar{B}$$

Dæmi

title	year	length	studioName	starName
Star Wars	1977	124	Fox	Carrie Fisher
Star Wars	1977	124	Fox	Harrison Ford
Star Wars	1977	124	Fox	Mark Hamill
Empire Strikes Back	1980	111	Fox	Harrison Ford
Terms of Endearment	1983	132	MGM	Debra Winger
Terms of Endearment	1983	132	MGM	Jack Nicholson
The Usual Suspects	1995	106	MGM	Kevin Spacey

Hvaða fallákveður gilda um þessi vensl?

Dæmi

Númer	Heiti	Nafn	Notendanafn	Námsleið	Hópur	Dagur	Tími
Töl303G	Gagnasafnsfr.	ABCD	abcd	Eðlisfræði	d2	Mið	15:50-17:20
Töl303G	Gagnasafnsfr.	BCDE	bcde	Stærðfræði	d3	Mið	8:20-9:50
Töl303G	Gagnasafnsfr.	BCDE	cdef	T-fræði	d4	Mið	8:20-9:50

Lyklar

$A_1, \ldots, A_n$ mynda **lykil** í venslum R ef

1. $\bar{A} \to B_i$ fyrir alla aðra dálka B_i
2. Ekkert hlutmengi í $\bar{A}$ ákvarðar alla aðra dálka.

Mengi af dálkum er **yfirlykill** ef það inniheldur lykil.

Reglur um fallákveður

Gegnvirkni: fyrir $R(A, B, C)$ gildir ef $A \to B$ og $B \to C$ þá gildir $A \to C$.

Skipting/sameining: eftirfarandi er jafngilt

$\bar{A} \to \bar{B}$ og $\bar{A} \to B_1 \wedge \ldots \wedge \bar{A} \to B_m$

þar sem $\bar{B} = \{B_1, \ldots, B_m\}$

Fáfengilegar fallákveður

Fallákveða er fáfengileg (trivial) ef

$\bar{A} \to \bar{B}$ og $\bar{B} \subseteq \bar{A}$.

Fallákveðan $\bar{A} \to \bar{B}$ er jafngild

$\bar{A} \to \bar{C}$

þar sem $\bar{C} = \bar{B} \setminus \bar{A}$

Lokun (Closure)

Ef $\bar{A}$ er mengi af eiginleikum og S fallákveður fyrir vensl R, þá er lokun $\bar{A}$ mengi eiginleika B þ.a. S leiðir til þess að $\bar{A} \to B_i$ fyrir öll $B_i \in \bar{B}$.

Lokunin er táknuð með $\bar{A}^+$.

Reiknirit fyrir lokun

Inntak, vensl R og fallákveður S og mengi $\bar{A}$

1. Skiptum upp fallákvðum í S þ.a. aðeins einn eiginleiki komi fyrir hægra megin.
2. Látum $X = \bar{A}$.

3. Leitum að fallákveðu $\bar{B} \to C$ þ.a. $\bar{B} \subseteq X$ en $C \notin X$. Bætum C við X og endurtökum.
4. skilum X.

Gegnvirkniregla

Útvíkkun á fyrri reglu, gildir líka fyrir mengi af dálkum, ekki bara staka dálka.

Ef $\bar{A} \to \bar{B}$ og $\bar{B} \to \bar{C}$ þá gildir $\bar{A} \to \bar{C}$.

T.d. fyrir einhvern "Movies" gagnagrunn gildir

$title, year \to studioName$ og $studioName \to studioAddress$ sem leiðir til þess að $title, year \to studioAddress$

Grunnur

Fyrir fallákveður S segjum við að ef S' er jafngilt S þá er S' **grunnur** (basis) fyrir S.

Grunnur B er lággrunur (minimal basis) ef

1. allar fallákveður í B eru með einn eiginleika hægra megin
2. ef fallákveða er tekin úr B þá er B ekki lengur grunnur
3. ef við tökum eiginleika úr vinstri hlið fallákveðu úr B þá er B ekki lengur grunnur

---

Fallákveður og vörpun

Ef R eru vensl með fallákveður S og $R_1 = \pi_L(R)$, hvaða fallákveður gilda um R_1

- leiða af S
- nota bara eiginleika í R_1

Reiknirit

Inntak, vensl R, fallákveður S og $R_1 = \pi_L(R)$

1. Látum $T = \emptyset$
2. Fyrir hvert hlutmengi X af eiginleikum í R_1 reiknum við X^+. Bætum við T öllum fallákveðum $X \to A$ þar sem $A \in X^+$ og $A \in L$.
3. T er grunnur fyrir fallákveður í R_1, finnum lággrunn
   - Ef ein fallákveða er afleiðing af hinum í T þá tökum við hana út
   - Látum $\bar{Y} \to B$ vera í T með a.m.k. tveimur breytum í $\bar{Y}$ og við tökum eina breytu út þ.a. eftir er $\bar{Z}$. Ef $\bar{Z} \to B$ er afleiðing af fallákveðunum í T þá skiptum við $\bar{Y} \to B$ út fyrir $\bar{Z} \to B$.
   - Endurtökum þar til ekkert breytist.

Hönnun gagnagrunna

Frávik (anomaly) er þegar gagnagrunnur er ekki alveg rétt hannaður

1. Endurtekningar: þegar gögn og n-dir eru endurteknar að óþörfu
2. Breytingar: við breytum gögnum á einum stað en gleymum að uppfæra sömu gögn á öðrum stað.
3. Eyðingar: þegar hlutar af gildi hverfa taka þeir aðra með sér

title	year	length	studioName	starName
Star Wars	1977	124	Fox	Carrie Fisher
Star Wars	1977	124	Fox	Harrison Ford
Star Wars	1977	124	Fox	Mark Hamill
Empire Strikes Back	1980	111	Fox	Harrison Ford
Terms of Endearment	1983	132	MGM	Debra Winger
Terms of Endearment	1983	132	MGM	Jack Nicholson
The Usual Suspects	1995	106	MGM	Kevin Spacey

Uppbrot á töflu (decomposition)

Ef R er tafla með dálka $\bar{A}$ þá getum við brotið R upp í töflur S og T með

1. $\bar{A} = \bar{B} \cup \bar{C}$
2. $S = \pi_B(R)$
3. $T = \pi_C(R)$

Sum uppbrot eru góð, t.d. $\{title, year, length, studionName\}$, $\{title, year, starName\}$

En önnur eru slæm $\{title, year\}$, $\{year, length, starName, studioName\}$.

BCNF (Boyce-Codd Normal Form)

BCNF lýsir því hvernig má brjóta upp töflur til að losna við frávik.

Vensl R eru á BCNF formi þþaa að ef $\bar{A} \rightarrow \bar{B}$ gildir (og eru ófáfengileg, þ.e. $\bar{B} \subsetneq \bar{A}$) þá er $\bar{A}$ yfirlykill.