TÖL303G

Gagnasafnsfræði

Snorri Agnarsson

Grunnur, þekja (basis, cover)

- Tvö söfn af fallákveðum, S og S' eru sögð **jafngild** ef allar fallákveður $\bar{A} \to B$ sem eru afleiðingar af S eru einnig afleiðingar af S' og öfugt
- Fyrir fallákveður (safn af fallákveðum) S' sem er jafngildar fallákveðum S segjum við einnig að S' sé **þekja eða grunnur** (cover, basis) fyrir S
- Grunnur B fyrir fallákveður S er **lágþekja eða lággrunnur** (minimal cover, minimal basis) ef
    1. Allar fallákveður í B hafa ein eigindi (eitt dálkanafn) hægra megin
    2. Ef fallákveða er fjarlægð úr B þá er B ekki lengur grunnur fyrir S
    3. Ef við fjarlægjum eiginleika úr vinstri hlið í einhverri fallákveðu í B þá er B ekki lengur grunnur fyrir S

Fallákveður og ofanvarp (FD‘s and projections)

- Látum R vera vensl með fallákveður S
- Látum venslin R_1 vera skilgreind með ofanvarpi, $R_1 = \pi_L(R)$
- Hvaða fallákveður gilda þá um R_1?
- Þær fallákveður eru afleiðingar af S
- Þær fallákveður nota einungis eigindi (dálka) í R_1

Algrím fyrir lágþekju (lággrunn, minimal basis, minimal cover) *hlutvensla*
Algorithm for minimal basis of subrelation (a projection of a relation)

Inntak: Vensl $\boldsymbol{R}$, fallákveður $\boldsymbol{S}$ og $\boldsymbol{R_1 = \pi_L(R)}$

Úttak: Lágþekja fyrir $\boldsymbol{R_1}$

1. Látum $T = \emptyset$
2. Fyrir sérhvert hlutmengi $\bar{X} \subseteq L$ af eigindum (dálkamengi) R_1 reiknum við $\bar{X}^+$
3. Bætum við T öllum fallákveðum $\bar{X} \rightarrow A$ þar sem $A \in \bar{X}^+ \cap L$
4. T er nú grunnur fyrir fallákveður R_1 -- finnum nú lágþekju
   - Ef ein fallákveða í T er afleiðing af hinum þá fjarlægjum við hana úr T
   - Ef til er fallákveða $\bar{Y} \rightarrow B$ í T þar sem $\bar{Y}$ inniheldur a.m.k. tvo dálka og til er ekki tómt $\bar{Z} \subset \bar{Y}$ þannig að $\bar{Y} \rightarrow B$ er afleiðing af $(\mathrm{T} - \{\bar{Y} \rightarrow B\}) \cup \{\bar{Z} \rightarrow B\}$, þá gefum við T nýja gildið $(T - \{\bar{Y} \rightarrow B\}) \cup \{\bar{Z} \rightarrow B\}$ --- (fjarlægjum sem sagt $\bar{Y} \rightarrow B$ og setjum $\bar{Z} \rightarrow B$ í staðinn)
   - Endurtökum þetta tvennt þar til ekkert breytist, skilum þá T með áorðnum breytingum

Dæmi um lágþekju (lággrunn) – Example of minimal cover (minimal basis)

- Gerum ráð fyrir heildarvenslum $R(A, B, C)$ með fallákveðum $\{AB \to C, C \to B\}$, ásamt hlutvenslum $R_1(B, C)$

  Assume a relation $R(A, B, C)$ with FD‘s $\{AB \to C, C \to B\}$, and a projected relation $R_1(B, C)$

- Eina fallákveðan sem verkar inni í R_1 er þá $C \to B$ og lágþekjan fyrir R_1 er því $\{C \to B\}$

  The only FD that works inside R_1 is then $C \to B$ and the minimal basis for R_1 is therefore $\{C \to B\}$

Dæmi um lágþekju (lággrunn) – Example of minimal cover (minimal basis)

- Gerum ráð fyrir heildarvenslum $R(A, B, C, D, E, F, G, H, I, J)$ með falllákveðum
  $$\{AB \to C, BD \to EF, AD \to GH, A \to I, H \to J\}$$
  ásamt hlutvenslum $R_1(A, D, G, H, I, J)$
- Fallákveðurnar sem verka inni í R_1 eru þá
  $$\{AD \to G, AD \to H, A \to I, H \to J\}$$
  og þær mynda lágþekju fyrir R_1

Frávik í gagnagrunnum – Database Anomalies

Frávik (anomaly, illbrigði) geta gerst þegar gagnagrunnur er ekki alveg rétt hannaður

1. Endurtekningar (**update anomaly**): Þegar sömu upplýsingar eru endurteknar og skrá þarf eða breyta sömu upplýsingum á fleiri en einum stað
2. Eyðingar (**deletion anomaly**): Þegar eytt er upplýsingum hverfa aðrar upplýsingar
3. Innsetningar (**insertion anomaly**): Þegar ekki er hægt að skrá upplýsingar

```
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 (þáttun, decomposition) á töflu Decomposition of a table (relation)

- Ef tafla (vensl) R hefur 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, til dæmis
  $$\{title, year, length, studioName\}, \{title, year, starName\}$$
- Önnur eru slæm, til dæmis
  $$\{title, year\}, \{year, length, starName, studioName\}$$

BCNF staðalsnið (Boyce-Codd Normal Form)

- BCNF segir til um hvernig skipuleggja má töflur til að losna við frávik (illvik, anomaly)
- Vensl R eru á BCNF sniði þá og því aðeins að
  - ef fallákveða $\bar{A} \rightarrow \bar{B}$ gildir innan R, og er ófáfengileg (þ.e. $\bar{B} \subseteq \bar{A}$ gildir ekki) þá er $\bar{A}$ yfirlykill
  - Með öðrum orðum, ef $\bar{A}$ ákvarðar meira en sjálft sig innan R þá ákvarðar $\bar{A}$ allt innan R
- Það eru til fleiri staðalsnið (1NF, 2NF, **3NF**, **BCNF**, EKNF, 4NF, 5NF, o.fl.)
  - við munum leggja áherslu á BCNF og 3NF

Sömu upplýsingar

- Þegar við þáttum (brjótum upp) vensl á réttan hátt, þá er hægt að fá upphaflegu venslin með náttúrlegri tengingu (natural join)
- Chase algrímið (3.4.2, bls. 92) notar þáttunina og fallákveður til að komast að því hvort gögnin varðveitist nákvæmlega (lossless join)
- Ef R er tafla og R er þáttuð í hlutvensl $R_1 = \pi_{S_1}(R), \ldots, R_n = \pi_{S_n}(R)$ þar sem $S_1, \ldots, S_n$ eru eigindamengin fyrir hin mismunandi hlutvensl, þá viljum við að þetta gildi:
$$R = R_1 \bowtie \cdots \bowtie R_n$$
- Chase algrímið (sjá síðar) staðfestir hvort:
$$R \supseteq R_1 \bowtie \cdots \bowtie R_n$$
- Ef svo er þá er þáttunin taplaus (lossless join) því öruggt er að:
$$R \subseteq R_1 \bowtie \cdots \bowtie R_n$$

Almenn markmið þáttunar (uppbrots) vensla

1. Viðhald allra upplýsinga (lossless join, taplausar tengingar)
   - Er $R = R_1 \bowtie \cdots \bowtie R_n$?
2. Viðhald fallákveða
   - Ef sanngildi fallákveðanna er tryggt í sérhverjum af hlutvenslunum $R_1, \dots, R_n$ hverjum fyrir sig, veldur það því að sanngildi fallákveðanna sé tryggt í $R_1 \bowtie \cdots \bowtie R_n$?
3. Útrýming frávika (anomalies)

- Það er ekki alltaf hægt að ná öllum markmiðum samtímis
  - Við verðum stundum að velja milli 2 og 3 – við fórnum ekki 1 og veljum líklega frekar 2 en 3

Dæmi um chase

- Íhugum $R(A, B, C)$ með fallákveðum $\{AB \to C, C \to B\}$
- Þáttum í $R_1(A, C), R_2(B, C)$ (fallákveðan $AB \to C$ er þá ekki innan neinnar töflu, en veldur það vandræðum? – chase gefur svarið)
- Notum chase algrím á þáttunina, hér er byrjunarstaðan:

A	B	C
a	b_1	c
a_2	b	c

- klárum á töflunni/skjánum í fyrirlestri

Annað dæmi um chase

- Íhugum $R(A, B, C)$ með falllákveðum $\{AB \rightarrow C, C \rightarrow B\}$
- Þáttum í $R_1(A, B), R_2(B, C)$ (falllákveðan $AB \rightarrow C$ er þá aftur ekki innan neinnar töflu, en veldur það nú vandræðum?)
- Notum chase algrím á þáttunina, hér er byrjunarstaðan:

A	B	C
a	b	c_1
a_2	b	c

- Við getum ekkert gert – engar raðir er hægt að tengja saman – þáttunin er **ekki taplaus**

Uppbrot (þáttun) yfir í BCNF

- Við getum tekið vensl R sem ekki eru á BCNF staðalsniði og þáttað þau í safn af nýjum venslum þannig að

1. Nýju venslin eru öll á BCNF staðalsniði
2. Hægt er að endurmynda gögnin úr R með tengingum
3. Gott markmið (sem ekki alltaf næst) er að til sé lágþekja (lággrunnur) þannig að sérhver fallákveða í lágþekjunni er innan einna þeirra vensla sem út koma úr þáttuninni

BCNF algrím

Inntak: Vensl $\boldsymbol{R}$, fallákveður $\boldsymbol{S}$
Úttak: Mengi hlutvensla $\boldsymbol{R}$ sem er BCNF þáttun $\boldsymbol{R}$ miðað við $\boldsymbol{S}$

1. Ef R er á BCNF sniði þá skilum við $\{R\}$
2. Annars finnum við einhver BCNF frávik, $\bar{X} \to Y$ innan S^+, þ.e. ófáfengilega (nontrivial) fallákveðu þannig að $\bar{X}$ er ekki yfirlykill (superkey), þ.e. $\bar{X}$ dugar ekki til að einkvæmt ákvarða röð í R
   - Reiknum $\bar{X}^+$
   - Setjum $R_1 = \pi_{\bar{X}^+}(R)$
   - Setjum $R_2 = \pi_L(R)$ þar sem L er sammengið af $\bar{X}$og þeim dálkum í R sem ekki eru í $\bar{X}^+$
3. Reiknum fallákveður fyrir venslin R_1 og R_2, köllum þær S_1 og S_2
4. Reiknum endurkvæmt BCNF fyrir R_1, S_1 og R_2, S_2, skilum sammenginu af niðurstöðunum

Tilgangur BCNF og staðalsniða almennt

- Losna við öll frávik
  - Endurtekningafrávik – margskráning sömu upplýsinga
  - Breytingafrávik – breyta þarf sömu upplýsingunum á mörgum stöðum
  - Eyðingafrávik – eyðing getur valdið eyðingu of mikilla upplýsinga
  - Innsetningafrávik – ekki er hægt að skrá þekktar upplýsingar
- Varðveita tengsl upplýsinga
  - Fallákveðurnar tengja saman upplýsingar

Einfalt dæmi um BCNF þáttun

- Gerum ráð fyrir venslum $R(A, B, C, D, E)$ með $AB \rightarrow C$ og $BC \rightarrow DE$
- Eini mögulegi lykillinn í R er AB
- Fallákveðan $BC \rightarrow D$ brýtur BCNF skilyrði því BC er ekki yfirlykill í R
- Brjótum því R upp í $R_1(B, C, D, E)$ og $R_2(A, B, C)$, sem uppfyllir BCNF skilyrði
- Lykillinn í R_1 er BC, lykillinn í R_2 er AB (sami og í R)

Virkar BCNF alltaf?

- Oftast, en ekki alltaf, sjá í wikipediu og kafla 3.4.4 í bókinni, bls. 96-97
- Til dæmis er **ekki til** BCNF staðalsnið fyrir vensl $R(A,B,C)$ með fallákveðum $\{AB \to C, C \to B\}$ sem varðveitir allar fallákveður
- Þátta má í $R_1(A,C), R_2(C,B)$ en þá er fallákveðan $AB \to C$ ekki innan neinnar töflu
- Þessi þáttun er taplaus (lossless decomposition), þ.e. $R = R_1 \bowtie R_2$, en gera þarf sérstakar ráðstafanir í gagnagrunninn til að tryggja fallákveðuna $AB \to C$
  - Tryggja þarf að ekki séu til **mismunandi** n-dir (a,c) og (a,c') í R_1 ásamt (c,b) og (c',b) í R_2
  - Ef svo væri þá væru n-dirnar (a,c,b) og (a,c',b) **báðar** í $R_1 \bowtie R_2$, sem væri þá í mótsögn við fallákveðuna $AB \to C$
  - Einnig þarf að tryggja að fyrir sérhverja n-d (a,c) í R_1 sé til n-d (c,b) R_2
- Hér veljum við frekar 3NF staðalsnið, sem gefur þáttunina $R_1(A,B,C)$, $R_2(C,B)$
  - Þægilegt að skorða með því að heimta að öll (b,c) í R_1 séu einnig í R_2 (tvískráning!)
  - Þáttunin kemur beint af augum úr lágþekjunni $\{AB \to C, C \to B\}$
  - Athugið samt að BCNF er almennt betra en 3NF (ekki ef fallákveður glatast)

Annað dæmi um chase

- Íhugum $R(A, B, C, D, E)$ með fallákveðum $\{AB \rightarrow C, BC \rightarrow DE\}$
- Þáttum í $R_1(B, C, D, E), R_2(A, B, C)$, sem uppfyllir BCNF án þess að glata fallákveðum
- Notum chase algrím á þáttunina, hér er byrjunarstaðan:

A	B	C	D	E
a_1	b	c	d	e
a	b	c	d_2	e_2

- klárum á töflunni í fyrirlestri

3NF staðalsnið

- 3NF er annað snið til að skipuleggja töflur til að losna við frávik
- Eilítið veikara en BCNF
  - Meira um tvískráningar upplýsinga
- Vensl R eru á 3NF sniði þá og því aðeins að
  - ef fallákveða $\bar{A} \rightarrow \bar{B}$ gildir innan R, og er ófáfengileg (þ.e. $\bar{B} \subseteq \bar{A}$ gildir ekki)
  - þá er annaðhvort $\bar{A}$ yfirlykill fyrir R **eða** sérhvert eigindi í $\bar{B} - \bar{A}$ er hluti af einhverjum mögulegum lykli

3NF algrím

Inntak: Vensl $\boldsymbol{R}$, fallákveður $\boldsymbol{F}$
Úttak: Mengi hlutvensla $\boldsymbol{R}$ sem er 3NF þáttun $\boldsymbol{R}$ miðað við $\boldsymbol{S}$

1. Finnum lágþekju (lággrunn) G sem er jafngild fallákveðusafninu F
2. Fyrir sérhverja fallákveðu $\bar{X} \rightarrow \bar{Y}$ í G búum við til hlutvensl $R_i(\bar{X}\bar{Y})$
   - Til dæmis ef $\bar{X} = ABC$ og $\bar{Y} = DE$ þá búum við til $R_i(A, B, C, D, E)$
3. Ef einhver venslanna $R_1, \dots, R_n$ sem út koma innihalda mögulegan lykil fyrir R (þ.e. eigindi venslanna mynda yfirlykil fyrir R) þá skilum við strax $\{R_1, \dots, R_n\}$
4. Annars búum við til ný hlutvensl $R'(\bar{X})$ þar sem $\bar{X}$er mögulegur lykill R og skilum síðan $\{R_1, \dots, R_n, R'\}$

Algrím fyrir lágþekju

Inntak: Safn F af falllákveðum
Úttak: Samsvarandi lágþekja G

1. Frumstillum $G = F$
2. Breytum ákveðum $X \to YZ$ í tvær eða fleiri ákveður $X \to Y$ og $X \to Z$
3. Breytum ákveðum $XY \to Z$ í $X \to Z$ ef G^+ breytist ekki
4. Fjarlægjum ákveður $X \to Z$ úr G ef G^+ breytist ekki

- Þegar þessu lýkur (ekkert meira er hægt að gera) er G lágþekja fyrir falllákveðusafnið F
- Athugið samt að stundum er fleiri en ein möguleg lágþekja

Einfalt dæmi um 3NF þáttun

- Gerum ráð fyrir venslum $R(A, B, C, D, E)$ með $AB \to C$ og $BC \to DE$
- Eini mögulegi lykillinn í R er AB
- Lágþekja er $\{AB \to C, BC \to DE\}$
- Brjótum því R upp í $R_1(B, C, D, E)$ og $R_2(A, B, C)$, sem uppfyllir 3NF skilyrði (og reyndar einnig BCNF skilyrði, sem er sterkara skilyrði)
- Lykillinn í R_1 er BC, lykillinn í R_2 er AB (sami og í R)
- Sama þáttun og við sáum áður sem BCNF
- Vorum búin að keyra chase og staðfesta taplausar tengingar

Frumeigindi, 3NF og BCNF

- **Skilgreining:** Gerum ráð fyrir venslum R með safni fallákveða F. Eigindi A í R eru þá **frumeigindi** í R ef til er mögulegur lykill sem inniheldur A.
- Munurinn á BCNF og 3NF er sá að í 3NF er **leyfilegt** að fallákveða $\bar{X} \rightarrow Y$ sé til staðar innan vensla R þótt $\bar{X}$ sé ekki yfirlykill R, ef Y er frumeigindi (m.v. að Y sé eitt eigindi)
- Í BCNF, hins vegar, er aðeins leyfilegt að fallákveða $\bar{X} \rightarrow Y$ sé til staðar innan vensla R ef $\bar{X}$ er yfirlykill R

Samband 3NF og BCNF

- Allar þáttanir sem eru BCNF eru 3NF
- Sumar þáttanir sem eru 3NF eru ekki BCNF
- Sem sagt: BCNF $\Rightarrow$ 3NF, en ekki öfugt

Yfirlit

- Lokun eigindamengis (dálkasafns)
- Lykill vensla
- Lágþekja (lággrunnur) fallákveðusafns
- Prófun taplausra tenginga
- Þáttun í BCNF og 3NF
- Þáttum helst í BCNF (til að minnka endurtekningar)
  - **með taplausum tengingum og viðhaldi fallákveða**
- Ef það bregst sættumst við á 3NF
  - **með taplausum tengingum og viðhaldi fallákveða**

Viðameiri dæmi

- Íhugum heildarvensl $R(A, B, C, E, F, G, H, I, J)$ með fallákveðum
$$AB \rightarrow C, BD \rightarrow EF, AD \rightarrow GH, A \rightarrow I, H \rightarrow J$$
- Finnum lykil fyrir R
- Finnum 3NF þáttun á R
- Finnum BCNF þáttun á R

Lykill

- Lykill verður að innihalda A, B og D
  - Vegna þess að þau eigindi koma hvergi fyrir hægra megin
- $ABD^+ = ABCDEFGHIJ$
- ABD er því (mögulegur) lykill

3NF þáttun

- Falllákveðusafnið
$$AB \rightarrow C, BD \rightarrow EF, AD \rightarrow GH, A \rightarrow I, H \rightarrow J$$
  er nú þegar lágþekja
- 3NF þáttun er því
$$R_1(A, B, C),$$
$$R_2(B, D, E, F),$$
$$R_3(A, D, G, H),$$
$$R_4(A, I),$$
$$R_5(H, J),$$
$$R_6(A, B, D)$$

BCNF þáttun

- $R(A, B, C, D, E, F, G, H, I, J)$
  Lykill: ABD. Þáttum á AB
  - $R_1(A, B, C, I)$
    Lykill: AB. Þáttum á A
    - $R_{11}(A, I)$
      Lykill: A. BCNF
    - $R_{12}(A, B, C)$
      Lykill: AB. BCNF
  - $R_2(A, B, D, E, F, G, H, J)$
    Lykill: ABD. Þáttum á BD
    - $R_{21}(B, D, E, F)$
      Lykill: BD. BCNF
    - $R_{22}(A, B, D, G, H, J)$
      Lykill: ABD. Þáttum á AD
      - $R_{221}(A, D, G, H, J)$
        Lykill: AD. Þáttum á H
        - $R_{2211}(H, J)$
          Lykill: H. BCNF
        - $R_{2212}(A, D, G, H)$
          Lykill: AD. BCNF
      - $R_{222}(A, B, D)$
        Lykill: ABD. BCNF

Falllákveður:

$AB \rightarrow C$

$BD \rightarrow EF$

$AD \rightarrow GH$

$A \rightarrow I$

$H \rightarrow J$

BCNF þáttun 2

Fallákveður:

$AB \to C$

$BD \to EF$

$AD \to GH$

$A \to I$

$H \to J$

- $R(A, B, C, D, E, F, G, H, I, J)$
  Lykill: ABD. Þáttum á AD
  - $R_1(A, D, G, H, I, J)$
    Lykill: AD. Þáttum á A
    - $R_{11}(A, I)$
      Lykill: A. BCNF
    - $R_{12}(A, D, G, H, J)$
      Lykill: AD. Þáttum á H
      - $R_{121}(H, J)$
        Lykill: H. BCNF
      - $R_{122}(A, D, G, H)$
        Lykill: AD. BCNF
  - $R_2(A, B, C, D, E, F)$
    Lykill: ABD. Þáttum á BD
    - $R_{21}(B, D, E, F)$
      Lykill: BD. BCNF
    - $R_{22}(A, B, C, D)$
      Lykill: ABD. Þáttum á AB
      - $R_{221}(A, B, C)$
        Lykill: AB. BCNF
      - $R_{222}(A, B, D)$
        Lykill: ABD. BCNF