Informationssysteme

Sommersemester 2016

Prof. Dr.-Ing. Sebastian Michel
TU Kaiserslautern

smichel@cs.uni-kl.de

Übersicht

Kommende ∼2 Vorlesungen

- Was ist ein guter relationaler Entwurf und wie kann ein schlechter Entwurf verbessert werden?

Danach betrachten wir, wie ein Datenbanksystem aufgebaut ist bzw. im Kern funktioniert, zuerst bzgl. Performance.

- Wie werden Daten auf Festplatte abgelegt?
- Wie werden (SQL) Anfragen ausgeführt und welche Kosten fallen dabei an?
- Wie kann effizient auf Daten zugegriffen werden?
- Wie kann die Ausführung einer Anfrage optimiert werden?

Danach kommen weitere Aspekte, wie Transaktionsverwaltung (=Fehlerbehandlung sowie Mehrbenutzersynchronisation).

Relationale Entwurfstheorie

ProfVorl						
PersNr	Name	Rang	Raum	VorlNr	Titel	SWS
2125	Sokrates	C4	226	5041	Ethik	4
2125	Sokrates	C4	226	5049	Mäeutik	2
2125	Sokrates	C4	226	4052	Logik	4
...	...	...	...	...	...	...
2132	Popper	C3	52	5259	Der Wiener Kreis	2
2137	Kant	C4	7	4630	Die 3 Kritiken	4

Wieso ist dies ein schlechtes Schema?

ProfVorl						
PersNr	Name	Rang	Raum	VorlNr	Titel	SWS
2125	Sokrates	C4	226	5041	Ethik	4
2125	Sokrates	C4	226	5049	Mäeutik	2
2125	Sokrates	C4	226	4052	Logik	4
...	...	...	...	...	...	...
2132	Popper	C3	52	5259	Der Wiener Kreis	2
2137	Kant	C4	7	4630	Die 3 Kritiken	4

Wieso ist dies ein schlechtes Schema?

- **Änderungsanomalien:** Sokrates zieht um, von Raum 226 in Raum 338. Was passiert?
- **Einfügeanomalien:** Neuer Professor ohne Vorlesungen?
- **Löschanomalien:** Letzte Vorlesung eines Profs wird gelöscht. Was passiert?

Funktionale Abhängigkeiten

Functional Dependency (FD)

- Schema $\mathcal{R} = \{A, B, C, D\}$
- Ausprägung R
- Seien $\alpha \subseteq \mathcal{R}$ und $\beta \subseteq \mathcal{R}$
- $\alpha \rightarrow \beta$ genau dann, wenn $\forall r, s \in R$ gilt: $r.\alpha = s.\alpha \Rightarrow r.\beta = s.\beta$
- D.h. **die α-Werte bestimmen die β-Werte funktional** (=eindeutig)

R			
A	B	C	D
a4	b2	c4	d3
a1	b1	c1	d1
a1	b1	c1	d2
a2	b2	c3	d2
a3	b2	c4	d3

$$\{A\} \rightarrow \{B\}$$

$$\{A\} \rightarrow \{C\}$$

$$\{C, D\} \rightarrow \{B\}$$

Aber nicht $\{B\} \rightarrow \{C\}$
Notation. auch ohne { bzw }:
$C,D \rightarrow B$ oder auch $CD \rightarrow B$

Einhaltung funktionaler Abhängigkeiten

Alternative Formulierung

- Die FD $\alpha \rightarrow \beta$ ist in R erfüllt, wenn für jede mögliche Ausprägung von α gilt:
$$|\pi_\beta(\sigma_{\alpha=c}(R))| \leq 1$$
- "die Menge aller Tupel von R mit $\alpha = c$ (für beliebiges c) projiziert auf β enthält keinen oder genau einen Wert"
- ansonsten wäre es auch keine "Funktion"

Daraus folgt ein einfacher Algorithmus:

boolean giltFD(Relation R, FD $\alpha \rightarrow \beta$){
    sortiere R nach α Werten
    für jede Gruppe $G_i \subseteq R$ von Tupeln mit Wert α_i aus α:
        falls nicht alle β_i aus β identisch sind:
            return falsch;
    return wahr;
}

Schlüssel

- $\alpha \subseteq \mathcal{R}$ ist ein Superschlüssel, falls gilt: $\alpha \rightarrow \mathcal{R}$
- trivial: es gilt $\mathcal{R} \rightarrow \mathcal{R}$
- Allerdings sind Superschlüssel nicht notwendigerweise minimal!

Volle funktionale Abhängigkeit:

- β ist voll funktional abhängig von α (Notation: $\stackrel{\bullet}{\rightarrow}$) genau dann wenn gilt:
  - $\alpha \rightarrow \beta$
  - α kann nicht mehr verkleinert (=linksreduziert) werden, d.h.
    - $\forall A \in \alpha$ folgt, dass $(\alpha - \{A\}) \rightarrow \beta$ nicht gilt, bzw. alternativ:
    - $\forall A \in \alpha : \neg((\alpha - \{A\}) \rightarrow \beta)$
- $\alpha \subseteq \mathcal{R}$ ist ein **Kandidatenschlüssel**, falls gilt: $\alpha \stackrel{\bullet}{\rightarrow} \mathcal{R}$

Ein Kandidatenschlüssel wird als **Primärschlüssel** ausgewählt!

Beispiel zur Schlüsselbestimmung

Städte			
Name	BLand	Vorwahl	EW
Frankfurt	Hessen	69	650000
Frankfurt	Brandenburg	335	84000
München	Bayern	89	1200000
Passau	Bayern	851	50000
Saarbrücken	Saarland	681	175000
Kaiserslautern	Rheinland-Pfalz	631	100000
...	...	...	...

Kandidatenschlüssel von Städte:

- {Name, BLand}
- {Name, Vorwahl}

Bestimmung funktionaler Abhängigkeiten

- Tabelle **Professoren**: $\{[\text{PersNr, Name, Rang, Raum, Ort, Straße, PLZ, Vorwahl, BLand, EW, Landesregierung}]\}$

- $\{\text{PersNr}\} \rightarrow \{\text{PersNr, Name, Rang, Raum, Ort, Straße, PLZ, Vorwahl, BLand, EW, Landesregierung}\}$
- $\{\text{Ort,BLand}\} \rightarrow \{\text{EW, Vorwahl}\}$
- $\{\text{PLZ}\} \rightarrow \{\text{BLand, Ort, EW}\}$
- $\{\text{BLand, Ort, Straße}\} \rightarrow \{\text{PLZ}\}$
- $\{\text{BLand}\} \rightarrow \{\text{Landesregierung}\}$
- $\{\text{Raum}\} \rightarrow \{\text{PersNr}\}$

Außerdem kann abgeleitet werden:

- $\{\text{Raum}\} \rightarrow \{\text{PersNr, Name, Rang, Raum, Ort, Straße, PLZ, Vorwahl, BLand, EW, Landesregierung}\}$
- $\{\text{PLZ}\} \rightarrow \{\text{Landesregierung}\}$

Herleitung weiterer FDs

- Aus einer Menge $\mathcal{F}$ von FDs sind weitere FDs herleitbar
- $\mathcal{F}^+$ wird **Hülle** (**closure**) von $\mathcal{F}$ genannt.
- $\mathcal{F}^+$ besteht aus allen aus $\mathcal{F}$ herleitbaren FDs
- Inferenzregeln, die **Armstrong Axiome**, beschreiben die Herleitung

Die Armstrong-Axiome

α,β, γ, δ sind Teilmengen der Attribute aus $\mathcal{R}$

- **Reflexivität:** Falls β eine Teilmenge von α ist ($\beta \subseteq \alpha$) dann gilt immer $\alpha \rightarrow \beta$.
  Insbesondere gilt also immer $\alpha \rightarrow \alpha$.
- **Verstärkung:** Falls $\alpha \rightarrow \beta$ gilt, dann gilt auch $\alpha\gamma \rightarrow \beta\gamma$. Hierbei stehe z.B. $\alpha\gamma$ für $\alpha \cup \gamma$
- **Transitivität:** Falls $\alpha \rightarrow \beta$ und $\beta \rightarrow \gamma$ gilt, dann gilt auch $\alpha \rightarrow \gamma$.

Die Armstrong-Axiome sind korrekt und vollständig.

D.h. Mit Hilfe dieser Axiome können alle gültigen FDs hergeleitet werden.

Die Armstrong-Axiome (2)

Nicht notwendig, aber oftmals komfortabel für Herleitungen:

- **Vereinigungsregel:**
  Wenn $\alpha \to \beta$ und $\alpha \to \gamma$ gelten, dann auch $\alpha \to \beta\gamma$.

- **Dekompositionsregel:**
  Wenn $\alpha \to \beta\gamma$ gilt, dann gelten auch $\alpha \to \beta$ und $\alpha \to \gamma$.

- **Pseudotransitivitätsregel:**
  Wenn $\alpha \to \beta$ und $\gamma\beta \to \delta$, dann gilt auch $\alpha\gamma \to \delta$

Attributhülle

Die Attributhülle AttrHülle($\mathcal{F}$,α) einer Attributmenge α bzgl. FDs $\mathcal{F}$ ist die Menge von Attributen α^+, für die gilt $\alpha \rightarrow \alpha^+$.

Gegeben:

- eine Menge $\mathcal{F}$ von FDs
- eine Menge von Attributen $\alpha \subseteq \mathcal{R}$

Algorithmus:

Erg := α;
while (Änderungen an Erg) **do**
  for each FD $\beta \rightarrow \gamma$ **in** $\mathcal{F}$ **do**
    if $\beta \subseteq Erg$ **then** $Erg := Erg \cup \gamma$;
$\alpha^+ := Erg$;

Beispielanwendung

Ziel:

Bestimme, ob κ einen Superschlüssel einer Relation $\mathcal{R}$ bzgl. der FDs in $\mathcal{F}$ bildet.

Lösung:

Durch Aufruf AttrHülle($\mathcal{F}$,κ) erhalten wir κ^+.
Falls $\kappa^+ = \mathcal{R}$: κ ist Superschlüssel von $\mathcal{R}$.

Äquivalente FD-Mengen

- Im Allgemeinen: **Es gibt viele unterschiedliche äquivalente Mengen von funktionalen Abhängigkeiten.**
- **Zwei Mengen $\mathcal{F}$ und $\mathcal{G}$ heißen äquivalent** (Notation: $\mathcal{F} \equiv \mathcal{G}$), **wenn ihre Hüllen gleich sind**, d.h. $\mathcal{F}^+ = \mathcal{G}^+$.
- **Bedeutung: die gleichen Mengen von FDs müssen herleitbar sein.**

Beobachtung:

- $\mathcal{F}^+$ kann sehr groß sein
- viele redundanten Abhängigkeiten
- in der Praxis unübersichtlich

Ziel: **kleinstmögliche** Menge $\mathcal{F}_c$ finden, so dass immer noch gilt: $\mathcal{F}_c^+ = \mathcal{F}^+$.

Kanonische Überdeckung $\mathcal{F}_c$

Folgende Eigenschaften müssen erfüllt sein:

1. $\mathcal{F}_c \equiv \mathcal{F}$, **d.h.,** $\mathcal{F}_c^+ = \mathcal{F}^+$
2. **In $\mathcal{F}_c$ existieren keine FDs $\alpha \to \beta$, bei denen α oder β überflüssige Attribute enthalten.** D.h. es muss gelten:
   (a) $\forall A \in \alpha : (\mathcal{F}_c - (\alpha \to \beta) \cup ((\alpha - A) \to \beta)) \not\equiv \mathcal{F}_c$
   (b) $\forall B \in \beta : (\mathcal{F}_c - (\alpha \to \beta) \cup (\alpha \to (\beta - B))) \not\equiv \mathcal{F}_c$
3. **Jede linke Seite einer funktionalen Abhängigkeit in $\mathcal{F}_c$ ist einzigartig.** Dies kann durch sukzessive Anwendung der Vereinigungsregel auf FDs der Art $\alpha \to \beta$ und $\alpha \to \gamma$ erzielt werden, so dass die beiden FDs durch $\alpha \to \beta\gamma$ ersetzt werden.

Algorithmus zur Bestimmung von $\mathcal{F}_c$

1. **Führe für jede FD $\alpha \rightarrow \beta \in \mathcal{F}$ die Linksreduktion durch**, also:
   - Überprüfe für alle $A \in \alpha$, ob A überflüssig ist, d.h. ob $\beta \subseteq AttrHülle(\mathcal{F}, \alpha - A)$
2. **Führe für jede (verbliebene) FD die Rechtsreduktion durch**:
   - Überprüfe für alle $B \in \beta$, ob $B \in AttrHülle(\mathcal{F} - (\alpha \rightarrow \beta) \cup (\alpha \rightarrow (\beta - B)), \alpha)$ gilt. In diesem Fall ist B auf der rechten Seite überflüssig und kann eliminiert werden, d.h. $\alpha \rightarrow \beta$ wird durch $\alpha \rightarrow (\beta - B)$ ersetzt.
3. **Entferne die FDs der Form $\alpha \rightarrow \emptyset$, die im 2. Schritt entstanden sind.**
4. **Fasse mittels Vereinigungsregel FDs der Form $\alpha \rightarrow \beta_1, ..., \alpha \rightarrow \beta_n$ zusammen**, so dass $\alpha \rightarrow (\beta_1 \cup ...\beta_n)$ verbleibt.

Beispiel für Herleitung einer kanonischen Überdeckung

- $\mathcal{F} = \{A \rightarrow B, B \rightarrow C, AB \rightarrow C\}$

- **Schritt 1:** Linksreduktion ersetzt $AB \rightarrow C$ durch $A \rightarrow C$
- **Schritt 2:** Rechtsreduktion ersetzt $A \rightarrow C$ durch $A \rightarrow \emptyset$
- **Schritt 3:** eliminiere $A \rightarrow \emptyset$
- **Schritt 4:** keine Zusammenfassungen notwendig

- Damit ist $\mathcal{F}_c = \{A \rightarrow B, B \rightarrow C\}$

Normalisierung von Relationen

Um Qualitätsprobleme im ursprünglichen Entwurf zu beheben, wird das bestehende Relationenschema $\mathcal{R}$ wird in mehrere Relationenschemata $\mathcal{R}_1, ..., \mathcal{R}_n$ zerlegt, die dann "besser" sind.

- Die Güte einer Zerlegung wird mit **Normalformen** beschrieben.
- Normalformen: 1NF, 2NF, 3NF, BCNF, 4NF, ...

Korrektheitskriterien für Zerlegung:

- **Verlustlosigkeit:** Die in der ursprünglichen Relationenausprägung R des Schemas $\mathcal{R}$ enthaltenen Daten müssen aus den Ausprägungen $R_1, ..., R_n$ der neuen Relationenschemata $\mathcal{R}_1, ..., \mathcal{R}_n$ rekonstruierbar sein.
- **Abhängigkeitsbewahrung:** Alle FDs in $\mathcal{F}_{\mathcal{R}}$ sollten in den $\mathcal{F}_{\mathcal{R}_1}, \mathcal{F}_{\mathcal{R}_2}, ...\mathcal{F}_{\mathcal{R}_n}$ bewahrt bleiben.

Verlustlosigkeit

- Zerlegung ist gültig, wenn: $\mathcal{R} = \mathcal{R}_1 \cup \mathcal{R}_2$
- D.h. alle Attribute aus $\mathcal{R}$ bleiben in der Zerlegung erhalten
- Wir definieren:
  - $R_1 := \pi_{\mathcal{R}_1}(R)$
  - $R_2 := \pi_{\mathcal{R}_2}(R)$
- Die Zerlegung von $\mathcal{R}$ in $\mathcal{R}_1$ und $\mathcal{R}_2$ ist **verlustlos**, falls für jede mögliche (gültige) Ausprägung R von $\mathcal{R}$ gilt:

$$R = R_1 \bowtie R_2$$

Die Zerlegung muss also durch einen natürlichen Verbund (Join) rekonstruierbar sein.

D.h. $\mathcal{R}_1 \cap \mathcal{R}_2 \neq \emptyset$ “sinnvoller” Schlüssel existiert