Introduction to Data and Knowledge Engineering Uebung: 4 - Vereinfachtes Syntheseverfahren

TECHNISCHE UNIVERSITÄT DARMSTADT

Aufgabe 4.4 Terminologie wichtiger Begriffe im Kontext des Relationenmodells

TECHNISCHE UNIVERSITÄT DARMSTADT

Begriff:	Informelle Beschreibung:
Attribut	Spalte einer Tabelle
Wertebereich	Mögliche Werte eines Attributes (auch Domäne)
Attributwert	Element eines Wertebereichs
Relationenschema	Menge von Attributen
Relation	Menge von Zeilen (Datensätze) einer Tabelle
Tupel	Zeile einer Tabelle
Datenbankschema	Menge von Relationenschemata
Datenbank	Menge von Relationen (Basisrelationen)
Schlüssel	Minimale Menge von Attributen, deren Werte eine Zeile einer Tabelle eindeutig identifizieren
Primärschlüssel	Ein beim Datenbankentwurf ausgezeichneter Schlüssel
Fremdschlüssel	Attributmenge, die in einer anderen Relation ein Schlüssel ist
Fremdschlüsselbedingung	Alle Attributwerte des Fremdschlüssels tauchen in der anderen Relation als Werte des Schlüssels auf
Abhängigkeitstreue	Alle gegebenen Abhängigkeiten sind durch Schlüssel repräsentiert
Verbundtreue	Die Originalrelationen können durch den Verbund der Basisrelationen wiedergewonnen werden

Aufgabe 4.5 Vereinfachtes Syntheseverfahren

TECHNISCHE UNIVERSITÄT DARMSTADT

Bestimmung aller möglichen Schlüsselkandidaten...

Zunächst betrachten wir die transitive Hülle (in der Literatur auch unter: "Attribut-Hülle" bekannt), um daraus herleiten zu können, welche Attribute als potentielle Schlüsselkandidaten in Frage kommen.

Linke FD-Seite:	**bestimmt:**	**Attribute:**					
AE	$\longrightarrow$	A	B		D	E	
B	$\longrightarrow$		B		D		
CD	$\longrightarrow$	A		C	D		
CE	$\longrightarrow$	A	B	C	D	E	F
BDF	$\longrightarrow$	A	B	C	D	E	F

Triviale Abhängigkeiten sind in blau hinterlegt.

Gegebene Abhängigkeiten sind in magenta hinterlegt.

Aufgabe 4.5 Vereinfachtes Syntheseverfahren

TECHNISCHE UNIVERSITÄT DARMSTADT

Es fällt auf das die Superschlüssel CE sowie BDF alle anderen Nichtschlüsselattributen funktional bestimmen. Da jedoch der Attribut D in BDF überflüssig ist, ergibt sich für die Menge der Schlüsselkandidaten $\mathcal{SK}$ insgesammt:

$$\mathcal{SK} = \{\{\mathtt{B}, \mathtt{F}\}, \{\mathtt{C}, \mathtt{E}\}\}$$

Zusatzfrage: Welcher dieser beiden Schlüsselkandidaten kann als Primärschlüssel ausgewählt werden?

Es stellt sich hier die jedoch Frage, warum D in BDF überflüssig ist? Hierfür müssen wir die Attributhülle von BF genauer betrachten...

Aufgabe 4.5 Vereinfachtes Syntheseverfahren

TECHNISCHE UNIVERSITÄT DARMSTADT

Die trivialen Abhängigkeiten sind: B und F, die gegebenen Abhängigkeiten lauten: C und E:

Linke FD-Seite:	bestimmt:	Attribute:					
BF	$\longrightarrow$		B	C		E	F

Über B erreichen wir D und können anschließend mit CD auch noch A erreichen, sodass insgesammt sämtliche Nichtschlüsselattribute von BF ohne Zuhilfenahme von D bestimmt werden können:

Linke FD-Seite:	bestimmt:	Attribute:					
BF	$\longrightarrow$	A	B	C	D	E	F

Damit kann D aus BDF entfernt werden.

Aufgabe 4.5 Vereinfachtes Syntheseverfahren

TECHNISCHE UNIVERSITÄT DARMSTADT

Die höchste vorliegende Normalform ist in diesem Fall: 1 NF.

Grund: D ist partiell von B abhängig.

Aufgabe 4.5 Vereinfachtes Syntheseverfahren

TECHNISCHE UNIVERSITÄT DARMSTADT

Dummy FD hinzufügen...

Linke FD-Seite:	**bestimmt:**	**Attribute:**					
ABCDEF	$\longrightarrow$	δ					
AE	$\longrightarrow$	A	B		D	E	
B	$\longrightarrow$		B		D		
CD	$\longrightarrow$	A		C	D		
CE	$\longrightarrow$	A	B	C	D	E	F
BDF	$\longrightarrow$	A	B	C	D	E	F

Frage: Wofür wird die Dummy-FD genau benötigt?

Aufgabe 4.5 Vereinfachtes Syntheseverfahren

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Die Hinzunahme einer Dummy-FD ist im Grunde ein Trick, um die Verbundtreue wiederzustellen, welche bei dem vereinfachten Syntheseverfahren auftreten kann. Ist die Synthese abgeschlossen, kann die Dummy-FD wieder entfernt werden.

Frage: Muss die Dummy-FD immer mitgeführt werden?

Prinzipiell schadet es nie die Dummy-FD zu verwenden. Falls diese überflüssig ist, wird diese mit den FD's, die eine äquivalente linke Seite haben zusammengefasst und führt damit nicht zu der Bildung einer neuen Relation.

Aufgabe 4.5 Vereinfachtes Syntheseverfahren

TECHNISCHE UNIVERSITÄT DARMSTADT

Entfernen überflüssiger Attribute...

- In `BDF` ist `D` überflüssig (wurde bereits gezeigt).
- In `ABCDEF` ist `ABDF` überflüssig.

Generell: Ein Attribut ist überflüssig, genau dann wenn man es aus der linken (bzw. rechten) Seite einer FD entfernen kann und die FD-Menge dabei äquivalent bleibt.

Aufgabe 4.5 Vereinfachtes Syntheseverfahren

TECHNISCHE UNIVERSITÄT DARMSTADT

Entfernen überflüssiger FD's...

Die einzige FD, die im gegebenen Szenario überflüssig ist lautet: $\texttt{AE} \longrightarrow \texttt{D}$.

Generell: Eine funktionale Abhängigkeit (FD) ist überflüssig, genau dann wenn man sie aus der Menge der FD's entfernen kann und FD-Menge dabei äquivalent bleibt.

Aufgabe 4.5 Vereinfachtes Syntheseverfahren

TECHNISCHE UNIVERSITÄT DARMSTADT

Bilden der Äquivalenzklassen...

Wir fassen in diesem Schritt alle FD's zu "Klassen" zusammen, sofern sie gleiche (bzw. äquivalente) linke Seiten besitzen. Jede Äquivalenzklasse ergibt anschliessend ein Relationenschema.

- $\mathtt{CE} \longrightarrow \delta$, $\mathtt{CE} \longrightarrow \mathtt{DF}$, $\mathtt{BF} \longrightarrow \mathtt{CE}$
- $\mathtt{AE} \longrightarrow \mathtt{B}$
- $\mathtt{B} \longrightarrow \mathtt{D}$
- $\mathtt{CD} \longrightarrow \mathtt{A}$

Aufgabe 4.5 Vereinfachtes Syntheseverfahren

TECHNISCHE UNIVERSITÄT DARMSTADT

Bilden der Realtionen-Schema R_i :

Links in jedem Tupel stehen alle enthaltenden Attribute pro Relation, rechts widerum stehen jeweils deren Primärschlüssel.

- $R_1 =$ (BCDEF, {{ B, F }, { C, E }})
- $R_3 =$ (ABE, {{ A, E }})
- $R_2 =$ (ACD, {{ C, D }})
- $R_4 =$ (DB, {{ B }})

Da D von FB partiell abhängig ist, ist hier die zweite Normalform immernoch verletzt. Dies jedoch lässt sich korrigieren, indem D aus R_1 entfernt wird.

Daraus erhalten wir: $R_1 =$ (BCEF, {{ B, F }, { C, E }})

Damit wäre die zweite Normalform erfüllt.
Ist diese jedoch gleichzeitig die höchste vorliegende NF?

Aufgabe 4.5 Vereinfachtes Syntheseverfahren

TECHNISCHE UNIVERSITÄT DARMSTADT

Nein, die 3NF ist hierdurch ebenfalls erfüllt!

Grund: Sämtliche Nichtschlüsselattribute sind in jeder Relation R_i nicht transitiv vom Primärschlüssel abhängig.

Damit ist die höchste vorliegende Normalform: 3NF.

Aufgabe 4.5 Vereinfachtes Syntheseverfahren

TECHNISCHE UNIVERSITÄT DARMSTADT

Frage: Was kann es für Folgen haben, falls eine Normalform verletzt wird?

Zunächst: es ist meistens schlecht, wenn ein Schema Relationen enthält, die eine Normalform verletzen (wobei es hier aber Ausnahmen und Kompromisse gibt). Wenn eine Normalform verletzt ist, werden Daten redundant gespeichert, und Informationen über verschiedene Konzepte vermischt, z.B.:

Vorlesungen an der TUD im Sommer-Semester 2010					
Vorlesung	**MHB Nummer**	**Prof.**	**Sprache**	**InternTel**	...
Web Mining	101	Fürnkranz	Deutsch	6238	...
Data Mining	168	Fürnkranz	Deutsch	6238	...
Künstl. Intelligenz	349	Fürnkranz	Deutsch	6238	...
Bildverarbeitung	155	Sakas	Deutsch	5153	...
IT Sicherheit	219	Katzenbeisser	Deutsch	5016	...
⋮	⋮	⋮	⋮	⋮	⋱

Aufgabe 4.5 Vereinfachtes Syntheseverfahren

TECHNISCHE UNIVERSITÄT DARMSTADT

Die Telefonnummer eines Dozenten (InternTel) wird, für jede von ihm gehaltene Vorlesung gespeichert.

Zwar stellt dies kein Problem dar, wenn eine Spalte einen Wert mehrfach enthält, aber im Beispiel gilt: Haben zwei Zeilen den gleichen Wert in der Spalte `Prof`, so müssen sie auch den gleichen Wert in der Spalte `InternTel` haben. Dies entspricht der funktionalen Abhängigkeit: `Prof` $\rightarrow$ `InternTel`.

Aufgrund dieser Bedingung ist einer der drei Einträge in der Spalte `InternTel` für Prof. "Fürnkranz" redundant.

Aufgabe 4.5 Vereinfachtes Syntheseverfahren

TECHNISCHE UNIVERSITÄT DARMSTADT

Redundante Information im konzeptionellen Schema sind also meistens schlecht:

- Speicherplatz wird verschwendet
- Doppelter Aufwand für Dateneingabe
- Wenn die Information aktualisiert wird, müssen auch alle redundanten Kopien aktualisiert werden. Vergißt man eine, so werden die Kopien inkonsistent (Update Anomalie).

Aufgabe 4.5 Vereinfachtes Syntheseverfahren

TECHNISCHE UNIVERSITÄT DARMSTADT

Frage: Wie hängen die Begriffe: "Aufspalten von Relationen", "Ausgangsrelation" sowie "Informationsverlust" miteinander?

Beim Aufspalten von Relationen ist es natürlich wichtig, dass die Transformation verlustlos verläuft, d.h. dass die Ausgangsrelation durch eine Anfrage wiederhergestellt werden kann, sodass kein Informationsverlust entsteht.