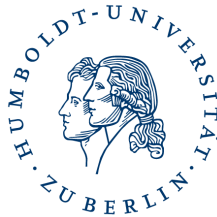


Einführung in die Datenbanktheorie

Vorlesung im Wintersemester

Prof. Dr. Nicole Schweikardt

Lehrstuhl Logik in der Informatik
Institut für Informatik
Humboldt-Universität zu Berlin



Version vom 11. Januar 2017

Inhaltsverzeichnis

1	Einleitung	5
1.1	Einführung ins Thema	5
1.2	Organisatorisches	11
1.3	Grundlegende Schreibweisen	11
2	Das Relationale Modell	13
2.1	Datenmodell	13
2.2	Anfragen	18
2.3	Datenkomplexität und kombinierte Komplexität	23
3	Konjunktive Anfragen	25
3.1	Deskriptiver Ansatz: regelbasiert, graphisch und logikbasiert	25
3.2	Auswertungskomplexität	45
3.3	Algebraischer Ansatz: SPC-Algebra und SPJR-Algebra	59
3.4	Homomorphismus-Satz, Statische Analyse und Anfrageminimierung	73
3.5	Azyklische Anfragen	88
3.6	Mengen-Semantik vs. Multimengen-Semantik	95
4	Datalog	101
4.1	Syntax, Semantik und Auswertungskomplexität	101
4.2	Grenzen der Ausdrucksstärke von Datalog	110
4.3	Datalog zur Simulation von Turingmaschinen	112

4.4	Statische Analyse	113
4.5	Einschränkung und Erweiterungen: nr-Datalog und Datalog mit Negation	116
5	Funktionale Abhängigkeiten	129
5.1	Notationen	129
5.2	The Chase — Die Verfolgungsjagd	132
5.3	Der Armstrong-Kalkül	138
6	Relationale Algebra	143
6.1	Definition und Beispiele	143
6.2	Anfrageauswertung und Heuristische Optimierung	147
7	Relationenkalkül	161
7.1	$CALC_{nat}$, $CALC_{adom}$ und $CALC_{di}$	161
7.2	Sicherer Relationenkalkül: $CALC_{sr}$	170
7.3	Statische Analyse und Auswertungskomplexität	176
7.4	Grenzen der Ausdrucksstärke	181
8	Zusammenfassung und Ausblick	185
8.1	Zusammenfassung	185
8.2	Ausblick auf weitere Themen	187

Kapitel 1

Einleitung

1.1 Einführung ins Thema

„Was“ statt „Wie“ — am Beispiel von „Tiramisu“

Folie 1

Tiramisu — Deklarativ

Aus Eigelb, Mascarpone und in Likör und Kaffee getränkten Biskuits hergestellte cremige Süßspeise

(aus: DUDEN, Fremdwörterbuch, 6. Auflage)

Tiramisu — Operationell

1/4 l Milch mit 2 EL Kakao und 2 EL Zucker aufkochen. 1/4 l starken Kaffee und 4 EL Amaretto dazugeben.

5 Eigelb mit 75 g Zucker weißschaumig rühren, dann 500 g Mascarpone dazumischen.

ca 200 g Löffelbiskuit.

Eine Lage Löffelbiskuit in eine Auflaufform legen, mit der Flüssigkeit tränken und mit der Creme überziehen. Dann wieder Löffelbiskuit darauflegen, mit der restlichen Flüssigkeit tränken und mit der restlichen Creme überziehen.

Über Nacht im Kühlschrank durchziehen lassen und vor dem Servieren mit Kakao bestäuben.

(aus: Gisela Schweikardt, handschriftliche Kochrezepte)

Folie 2

Der große Traum der Informatik

Imperative Vorgehensweise:

Beschreibung, wie das gewünschte Ergebnis erzeugt wird „Wie“

Deklarative Vorgehensweise:

Beschreibung der Eigenschaften des gewünschten Ergebnisses „Was“

Traum der Informatik:

Möglichst wenig „wie“, möglichst viel „was“

D.h.: Automatische Generierung eines Ergebnisses aus seiner Spezifikation

Realität:

Software-Entwicklung: Generierungs-Tools

Programmiersprachen: Logik-Programmierung, insbes. Prolog

ABER: Imperativer Ansatz überwiegt in der Praxis

Datenbanken: *Deklarative Anfragesprache ist Industriestandard! (SQL)*

Folie 3

Datenbanksysteme

Datenbank (DB)

- zu speichernde Daten
- Beschreibung der gespeicherten Daten (Metadaten)

Datenbankmanagementsystem (DBMS)

- Softwarekomponente zum Zugriff auf die Datenbank
- Eigenschaften / Kennzeichen:
 - *Sichere* Verwaltung von Daten: langlebig, große Menge von Daten ... im Sekundärspeicher
 - *Effizienter* Zugriff auf (große) Datenmengen in der DB

Datenbanksystem (DBS)

- DB + DBMS

Folie 4

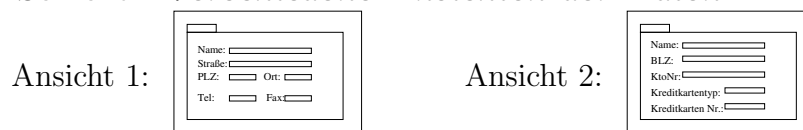
Wünschenswerte Eigenschaften eines DBS

- *Unterstützung eines Datenmodells:* „Darstellung“ der Daten für den Zugriff
- *Bereitstellung einer DB-Sprache:*
 - zur Datendefinition: Data Definition Language (DDL)
 - zur Datenmanipulation und zum Datenzugriff: Data Manipulation Language (DML)
- *Zugangskontrolle:* Wer darf wann auf welche Daten zugreifen bzw. verändern?
- *Datenintegrität:* Wahrung der Datenkonsistenz und -korrektheit
- *Robustheit:* Wahrung eines konsistenten Zustands der DB trotz ...
 - Datenverlusts bei Systemfehlern (CPU Fehler, Plattencrash)
 - fehlerhafter Beendigung eines DB-Programms oder einer DB-Interaktion
 - Verletzung der Datenintegrität oder von Zugriffsrechten
- *Zugriffskoordination bei mehreren DB-Benutzern:* Synchronisation, korrekter Zugriff, korrektes Ergebnis bzw. korrekter DB-Zustand
- *Effizienter Datenzugriff und Datenmanipulation:* schnelle Bearbeitung der Benutzeranfragen

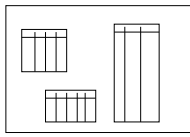
Folie 5

3-Schichten-Modell

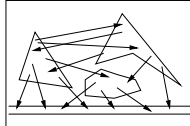
Externe Schicht: *Verschiedene Ansichten der Daten*



Logische Schicht: *Daten = Tabellen*



Physische Schicht: *Datenstrukturen, Speicherorganisation*



Folie 6

Anfragesprachen

Wünschenswerte Eigenschaften:

- *Möglichst viel „Was“*
Beschreiben der Eigenschaften des gewünschten Ergebnisses (deklarativ)
- *Möglichst wenig „Wie“*
Beschreiben, wie das gewünschte Ergebnis erzeugt werden soll (operationell)
- *Möglichst unabhängig von den Details der Datenorganisation*
Bezug auf logische Schicht oder externe Schicht, nicht auf physische Schicht

Der Preis der Bequemlichkeit und Unabhängigkeit:

- deklarative Anfragen verschieben die Arbeit vom Benutzer zum System
- System muss Anfrage in eine Folge von Operationen umwandeln
~> *Gefahr der Ineffizienz*
~> *Geht das überhaupt? Was ist die Auswertungskomplexität?*
- Andererseits: System hat große Freiheit in der Umsetzung, da kein Lösungsweg vorgeschrieben ist
~> *Potenzial für Optimierung*

Folie 7

Hauptthema dieser Vorlesung: Anfragesprachen

Typische Fragestellungen für diese Vorlesung:

- Wie lassen sich deklarative Anfragen in ausführbare Operationen umsetzen?

↪ *Äquivalenz von „Kalkül“ und „Algebra“*

- Welche Anfragen können in einer Anfragesprache gestellt werden, welche nicht?

↪ *Ausdrucksstärke von Anfragesprachen*

- Wie aufwändig ist die Auswertung von Anfragen prinzipiell?

↪ *Auswertungskomplexität*

- Wie lässt sich eine gegebene Anfrage möglichst effizient auswerten?

↪ *Anfrageoptimierung, statische Analyse (Erfüllbarkeit, Äquivalenz, ...)*

Folie 8

Inhaltsübersicht

1. *Einleitung*

2. *Das Relationale Modell*

- Datenmodell
- Anfragen
- Datenkomplexität und kombinierte Komplexität

3. *Konjunktive Anfragen*

- Deskriptiver Ansatz: regelbasiert, graphisch und logikbasiert
- Auswertungskomplexität
- Algebraischer Ansatz: SPC-Algebra und SPJR-Algebra
- Anfrageminimierung, statische Analyse und der Homomorphismus-Satz
- Azyklische Anfragen

- Mengen-Semantik vs. Multimengen-Semantik

4. *Anfragesprachen mit Rekursion — Datalog*

- Syntax und Semantik
- Auswertung von Datalog-Anfragen, Statische Analyse
- Datalog mit Negation

Folie 9

5. *Relationale Algebra*

- Definition und Beispiele
- Anfrageauswertung und Heuristische Optimierung
- Das Semijoin-Fragment der Relationalen Algebra

6. *Relationenkalkül*

- Syntax und Semantik
- Bereichsunabhängige Anfragen
- Äquivalenz zur Relationalen Algebra
- Auswertungskomplexität
- Statische Analyse

7. *Funktionale Abhängigkeiten*

- „The Chase“
- Anfrage-Optimierung unter Berücksichtigung funktionaler Abhängigkeiten
- der Armstrong-Kalkül

Folie 10

Literatur

- [AHV] Abiteboul, Hull, Vianu: Foundations of Databases, Addison-Wesley, 1995
- [M] Maier: *The Theory of Relational Databases*, Computer Science Press, 1983
- [AD] Atzeni, de Antonellis: *Relational Database Theory*, Benjamin Cummings, 1992
- [SSS] Schweikardt, Schwentick, Segoufin: Database Theory: Query Languages. *Kapitel 19 in Algorithms and Theory of Computation Handbook, 2nd Edition, Volume 2: Special Topics and Techniques*, Mikhail J. Atallah and Marina Blanton (editors), CRC Press, 2009.

1.2 Organisatorisches

Folie 11

Organisatorisches

- Webseite der Vorlesung:
www2.informatik.hu-berlin.de/logik/lehre/

1.3 Grundlegende Schreibweisen

Folie 12

- $\mathbb{N} := \{0, 1, 2, 3, \dots\}$
- $\mathbb{N}_{\geq 1} := \mathbb{N} \setminus \{0\}$
- Für $n \in \mathbb{N}_{\geq 1}$ ist $[n] := \{1, \dots, n\} = \{x \in \mathbb{N} : 1 \leq x \leq n\}$.
- Die Potenzmenge einer Menge M bezeichnen wir mit $\mathcal{P}(M)$.
D.h.: $\mathcal{P}(M) = \{X : X \subseteq M\}$.
- Ist M eine Menge, so schreiben wir $X \subseteq_e M$ um auszudrücken, dass X eine *endliche* Teilmenge von M ist.
- Wir benutzen folgende Abkürzungen:
 - „f.a.“ steht für „für alle“

- „ex.“ steht für „es existiert“ bzw. „es gibt“
- „s.d.“ steht für „so dass“

Kapitel 2

Das Relationale Modell

2.1 Datenmodell

Folie 13

Datenmodell

Der Begriff „Datenmodell“ umfasst:

- einen Rahmen zur Repräsentation bzw. Speicherung von Daten
- Operationen zum Zugriff auf Daten
- Mechanismen zur Beschreibung von erwünschten Eigenschaften (Integritätsbedingungen)

Der Begriff „Datenmodell“ ist nicht präzise definiert (im mathematischen Sinn).

Im Folgenden wird eine präzise Definition des „Relationalen Modells“ gegeben.

Folie 14

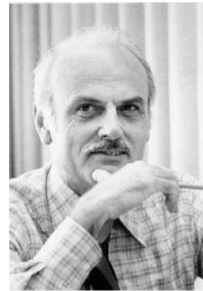
Das Relationale Modell

- Daten werden in Relationen („Tabellen“) organisiert
- Mengen-orientierte Operationen
- deklarative Anfragespezifikation

- effiziente Anfragebearbeitung
- 1970 eingeführt von Edgar F. Codd
- seit Ende der 80er Jahre „Industriestandard“

Beispiel-Relation:

A	B	C	D
a	1	z	9
b	2	z	8
c	1	y	8
d	4	x	7



Edgar F. Codd (1923-2003)

Folie 15

Grundbegriff: Relationsschema

Wir legen ein für alle Mal fest:

- Eine abzählbar unendliche Mengen **att** von *Attributnamen*. Diese Menge sei *geordnet* via $\leq_{\mathbf{att}}$.
- Eine abzählbar unendliche Menge **dom** von „potentiellen Datenbankeinträgen“ („Konstanten“) (**dom** steht für „Domäne“ bzw. „Domain“)
- Eine abzählbar unendliche Menge **rel** von *Relationsnamen*. Die Mengen **att**, **dom**, **rel** seien disjunkt.
- Eine Funktion $sorte : \mathbf{rel} \rightarrow \mathcal{P}_e(\mathbf{att})$, die jedem Relationsnamen eine endliche Menge von Attributnamen zuordnet, und zwar so, dass für alle $U \in \mathcal{P}_e(\mathbf{att})$ gilt: $sorte^{-1}(U)$ ist unendlich.
Notation: $\mathcal{P}_e(M)$ ist die Menge aller endlichen Teilmengen von M .
D.h.: Für jede endliche Menge U von Attributnamen gibt es unendlich viele verschiedene Relationsnamen R der Sorte U .
- Die *Stelligkeit* eines Relationsnamens R ist $ar(R) := |sorte(R)|$.
- Ein *Relationsschema* ist einfach ein Relationsname R .
- Manchmal schreiben wir kurz $R[U]$ für $sorte(R) = U$ und $R[k]$ für $ar(R) = k$.

Beispiel.

A	B	C	D
a	1	z	9
b	2	z	8
c	1	y	8
d	4	x	7

ist eine Relation vom Schema R mit $sorte(R) = \{A, B, C, D\}$ und $ar(R) = 4$.

Folie 16

Relationenschema vs. Relation

Relation $\hat{=}$ Tabelle

Tupel $\hat{=}$ Zeile in der Tabelle

Schreibweise: $t.A$ an Stelle von $t(A)$
für „Eintrag in Zeile t und Spalte A “

Beachte: *Mengensemantik*, d.h.:
Relation $\hat{=}$ die Menge aller Tabellenzeilen

Definition 2.1. Sei R ein Relationenschema.

- Ein R -Tupel ist eine Abbildung $t : sorte(R) \rightarrow \mathbf{dom}$.
- Eine R -Relation ist eine endliche Menge von R -Tupeln.
- $inst(R)$ bezeichnet die Menge aller R -Relationen.
(*inst* steht für „Instanzen“ bzw. „instances“)

Folie 17

Grundbegriffe: Datenbankschema und Datenbank

Definition 2.2.

- Ein *Datenbankschema* (kurz: *DB-Schema*) \mathbf{S} ist eine endliche, nicht-leere Menge von Relationenschemata.

Manchmal schreiben wir $\mathbf{S} = \{ R_1[U_1], \dots, R_n[U_n] \}$ um die Relationenschemata anzugeben, die zu \mathbf{S} gehören.

- Eine *Datenbank* (bzw. *Datenbankinstanz*) **I** vom Schema **S** ist eine Funktion, die jedem Relationsschema $R \in \mathbf{S}$ eine R -Relation zuordnet.
- $inst(\mathbf{S})$ bezeichnet die Menge aller Datenbanken vom Schema **S**.
- $schema(\mathbf{I}) := \mathbf{S}$ bezeichnet das Schema der Datenbank **I**.

Folie 18

Beispieldatenbank mit Kinodaten

Zur Illustration von Anfragen verwenden wir eine kleine Datenbank mit Kinodaten, bestehend aus

- einer Relation *Kinos*, die Informationen über Kinos (Name, Adresse, Stadtteil, Telefon) enthält.
- einer Relation *Filme*, die Informationen über Filme enthält (Titel, Regie, Schauspieler)
- einer Relation *Programm*, die Informationen zum aktuellen Kinoprogramm enthält (Kino, Titel, Zeit)

Folie 19

<i>Kinos</i>			
Name	Adresse	Stadtteil	Telefon
Babylon	Dresdner Str. 126	Kreuzberg	030 61 60 96 93
Casablanca	Friedenstr. 12-13	Adlershof	030 67 75 75 2
Filmtheater am Friedrichshain	Bötzowstr. 1-5	Prenzlauer Berg	030 42 84 51 88
Kino International	Karl-Marx-Allee 33	Mitte	030 24 75 60 11
Movimento	Kotbusser Damm 22	Kreuzberg	030 692 47 85
Urania	An der Urania 17	Schöneberg	030 21 89 09 1

<i>Filme</i>		
Titel	Regie	Schauspieler
Alien	Ridley Scott	Sigourney Weaver
Blade Runner	Ridley Scott	Harrison Ford
Blade Runner	Ridley Scott	Sean Young
Brazil	Terry Gilliam	Jonathan Pryce
Brazil	Terry Gilliam	Kim Greist
Casablanca	Michael Curtiz	Humphrey Bogart
Casablanca	Michael Curtiz	Ingrid Bergmann
Gravity	Alfonso Cuaron	Sandra Bullock
Gravity	Alfonso Cuaron	George Clooney
Monuments Men	George Clooney	George Clooney
Monuments Men	George Clooney	Matt Damon
Resident Evil	Paul Anderson	Milla Jovovich
Terminator	James Cameron	Arnold Schwarzenegger
Terminator	James Cameron	Linda Hamilton
Terminator	James Cameron	Michael Biehn
...

<i>Programm</i>		
Kino	Titel	Zeit
Babylon	Casablanca	17:30
Babylon	Gravity	20:15
Casablanca	Blade Runner	15:30
Casablanca	Alien	18:15
Casablanca	Blade Runner	20:30
Casablanca	Resident Evil	20:30
Filmtheater am Friedrichshain	Resident Evil	20:00
Filmtheater am Friedrichshain	Resident Evil	21:30
Filmtheater am Friedrichshain	Resident Evil	23:00
Kino International	Casablanca	18:00
Kino International	Brazil	20:00
Kino International	Brazil	22:00
Movimento	Gravity	17:00
Movimento	Gravity	19:30
Movimento	Alien	22:00
Urania	Monuments Men	17:00
Urania	Monuments Men	20:00

Datenbankschema der Kinodatenbank:

- Datenbankschema **KINO** = {*Kinos, Filme, Programm*}
- $sorte(Kinos) = \{Name, Adresse, Stadtteil, Telefon\}$
- $sorte(Filme) = \{Titel, Regie, Schauspieler\}$
- $sorte(Programm) = \{Kino, Titel, Zeit\}$

Wir schreiben **I_{KINO}**, um unsere konkrete Datenbank vom Schema **KINO** zu bezeichnen. Analog schreiben wir I_{Filme} , I_{Kinos} und $I_{Programm}$ für die konkreten Relationen, die zur Datenbank **I_{KINO}** gehören.

Attribute: Benannte vs. Unbenannte Perspektive

Sind die Attributnamen Teil des expliziten Datenbankschemas?

In SQL: ja! Beispiel:

```
SELECT Titel FROM Filme WHERE Schauspieler = 'Sigourney Weaver'
```

Aber werden die Namen vom System nicht „weg-compiliert“?

- *Benannte Perspektive:*
Ein Tupel über Relationenschema $R[U]$ ist eine Abbildung von U nach **dom**.
Schreibweise: $t = (A : a, B : 1, C : z, D : 9)$

- *Unbenannte Perspektive:*
Ein Tupel über Relationsschema $R[k]$ ist ein Element aus \mathbf{dom}^k
(Kartesisches Produkt aus k Kopien von \mathbf{dom}).
Schreibweise: $t = (a, 1, z, 9)$

Folie 23

Wir nutzen folgende Schreibweisen für

Konstanten (d.h. Elemente aus \mathbf{dom})	$a, b, c, \text{“Ingrid Bergmann”}, \dots$
Attributnamen	A, B, C, \dots
Mengen von Attributnamen	U, V, \dots
Relationsnamen (bzw. -schemata)	$R, R', R[U], R'[V], \dots$
Datenbankschemata	\mathbf{S}, \mathbf{S}'
Tupel	t, t', s
Relationen (d.h. Relations-Instanzen)	I, J
Datenbanken (Datenbank-Instanzen) .	\mathbf{I}, \mathbf{J}

2.2 Anfragen

Folie 24

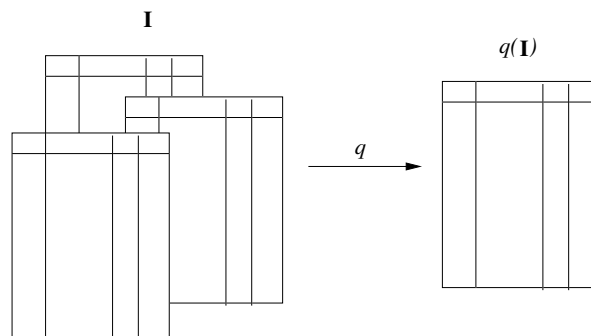
Beispiel-Anfragen

- (1) Wer führt Regie in “Blade Runner”?
- (2) Wo läuft “Gravity”?
- (3) Gib Adresse und Telefonnummer des “Kino International” aus!
- (4) Welche Kinos spielen einen Film mit “Matt Damon”?
- (5) Läuft zur Zeit ein Film von “James Cameron”?
- (6) Welche (je 2) Schauspieler haben schon in Filmen gespielt, in denen der jeweils andere Regie geführt hat?
- (7) Welche Regisseure haben in einem ihrer eigenen Filme mitgespielt?

- (8) Gib die 2-Tupel von Schauspielern an, die gemeinsam in einem Film gespielt haben!
- (9) Egal für welche Datenbank, gib als Antwort das Tupel (“*Terminator*”, “*Linda Hamilton*”) aus!
- (10) Wo wird “*Alien*” oder “*Brazil*” gespielt?
- (11) Welche Filme laufen in mindestens 2 Kinos?
- (12) In welchen Filmen spielt “*George Clooney*” mit oder führt Regie?
- (13) Gib alle Filme aus, die im “*Movimiento*” laufen oder in denen “*Sandra Bullock*” mitspielt!
- (14) Liste alle Schauspieler und den Regisseur von “*Monuments Men*” auf!
- (15) Welche Filme laufen nur zu 1 Uhrzeit?
- (16) In welchen Filmen spielt “*George Clooney*” mit, ohne Regie zu führen?
- (17) Welche Filme haben nur Schauspieler, die schon mal in einem Film von “*James Cameron*” mitgespielt haben?

Folie 25

Anfragen und Anfragefunktionen



Definition 2.3.

- Eine *Anfragefunktion* ist eine Abbildung q , die, für ein Datenbankschema S und ein Relationsschema R , jeder Datenbank I vom Schema S , eine Relation, $q(I)$ vom Schema R zuordnet. (q steht für „query“)

- Eine *Anfrage* ist eine Zeichenkette, die eine Anfragefunktion in einer bestimmten Syntax beschreibt.

Folie 26

Wünschenswerte Eigenschaften von Anfragefunktionen

Forderungen an eine Anfragefunktion q :

- (a) Das Ergebnis sollte nicht von Details der Speicherung abhängen, sondern nur von der logischen Sicht auf die Daten

Dies wird dadurch gewährleistet, dass q eine Funktion $inst(\mathbf{S}) \rightarrow inst(R)$ ist, für ein DB-Schema \mathbf{S} und ein Rel.schema R

- (b) Sie sollte berechenbar sein.

D.h. es sollte einen Algorithmus geben, der bei Eingabe einer Datenbank \mathbf{I} das Anfrageergebnis $q(\mathbf{I})$ berechnet.

- (c) Das Ergebnis sollte möglichst wenig von den einzelnen Datenwerten und möglichst viel von den Beziehungen der Daten untereinander abhängen

... siehe nächste Folie; Stichwort: „generisch“

Folie 27

Erläuterungen zu Eigenschaft (c)

Beispiel-Anfrage:

- (3) Gib Adresse und Telefonnummer des “Kino International” aus!

Eigenschaft (c):

- Wenn sich die Telefonnummer vom “Kino International” in der DB ändert, soll sich das Ergebnis der Anfrage entsprechend ändern.
- Aber wenn der Name “Kino International” sich in der DB ändert, soll das Ergebnis leer sein.
- *Allgemein:*

Werden Elemente der Datenmenge **dom**, die in der Anfrage nicht explizit als “Konstanten” vorkommen, umbenannt, so sollen sie im Ergebnis auch umbenannt werden.

- *Mathematische Präzisierung:* Begriff der *generischen Anfragefunktion*

Folie 28

Generische Anfragefunktionen

Definition 2.4.

Sei C eine endliche Menge von Datenwerten (kurz: $C \subseteq_e \mathbf{dom}$).

Eine Anfragefunktion q heißt C -*generisch*, falls für jede Datenbank \mathbf{I} (vom zu q passenden DB-Schema) und jede Permutation π von **dom** mit $\pi|_C = id$ (d.h. $\pi(c) = c$ für alle $c \in C$) gilt:

$$q(\pi(\mathbf{I})) = \pi(q(\mathbf{I})).$$

q heißt *generisch*, falls q \emptyset -generisch ist.

Illustration:

$$\begin{array}{ccc} \mathbf{I} & \xrightarrow{q} & q(\mathbf{I}) \\ \pi \downarrow & & \pi \downarrow \\ \pi(\mathbf{I}) & \xrightarrow{q} & q(\pi(\mathbf{I})) \end{array}$$

Beispiel:

(3) Gib Adresse und Telefonnummer des “Kino International” aus!
ist {“Kino International”}-generisch.

(7) Welche Regisseure haben in einem ihrer eigenen Filme mitgespielt?
ist generisch.

Folie 29

Boolesche Anfragen

Manche Anfragen lassen sich nur mit „ja“ oder „nein“ beantworten.

Beispiel: (5) Lläuft zur Zeit ein Film von “James Cameron”?

Konvention:

- Das Ergebnis ist eine 0-stellige Relation.
- Davon gibt es genau zwei Stück: \emptyset und $\{()\}$.
 $\{()\}$ steht für das „Tupel der Stelligkeit 0“.
- Vereinbarung:
 - \emptyset steht für „nein“
 - $\{()\}$ steht für „ja“

Folie 30

Anfragesprachen

Dieselbe Anfragefunktion kann in verschiedenen Anfragesprachen beschrieben werden.

Beispiel: (4) Welche Kinos spielen einen Film mit “Matt Damon”?

- *SQL:*

```
SELECT  Kinos.Name, Kinos.Adresse
FROM    Filme, Programm, Kinos
WHERE   Filme.Schauspieler = ‘‘Matt Damon’’ AND
        Filme.Titel = Programm.Titel AND
        Programm.Kino = Kinos.Name
```
- *Regelbasierte Anfrage:*

$$Ans(x_{Kino}, x_{Adr}) \leftarrow \begin{array}{l} Filme(x_{Titel}, x_{Regie}, \text{“Matt Damon”}), \\ Programm(x_{Kino}, x_{Titel}, x_{Zeit}), \\ Kinos(x_{Kino}, x_{Adr}, x_{St}, x_{Tel}) \end{array}$$
- *Relationenkalkül:*

$$\left\{ (x_{Kino}, x_{Adr}) : \exists x_T \exists x_R \exists x_Z \exists x_{St} \exists x_{Tel} (Filme(x_T, x_R, \text{“Matt Damon”}) \wedge \right. \\ \left. Programm(x_{Kino}, x_T, x_Z) \wedge Kinos(x_{Kino}, x_{Adr}, x_{St}, x_{Tel})) \right\}$$
- *Relationale Algebra:*

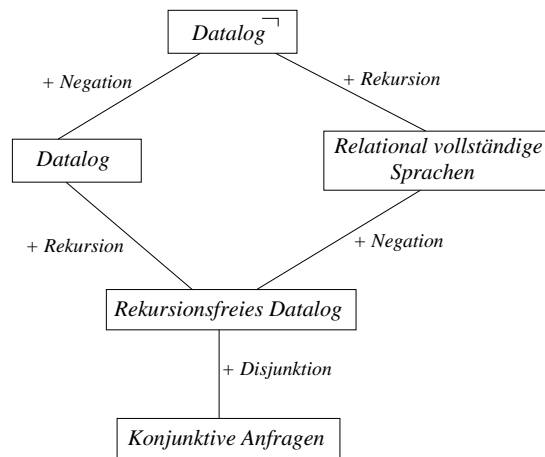
$$\pi_{Kino, Adresse} \left(\sigma_{\substack{Schauspieler = \\ \text{“Matt Damon”}}} (Filme \bowtie Programm \bowtie \delta_{Name \mapsto Kino}(Kinos)) \right)$$

Folie 31

Hierarchie der Anfragesprachen

Bemerkung: Anfragesprachen unterscheiden sich in ihrer Ausdrucksstärke.

Übersicht:



2.3 Datenkomplexität und kombinierte Komplexität

Folie 32

Typische Problemstellungen bzgl. Anfrageauswertung

Sei \mathcal{A} eine Anfragesprache.

Für eine Anfrage $Q \in \mathcal{A}$ schreiben wir $\llbracket Q \rrbracket$, um die von Q beschriebene *Anfragefunktion* zu bezeichnen.

AUSWERTUNGSPROBLEM FÜR \mathcal{A} :

Eingabe: Anfrage $Q \in \mathcal{A}$, Datenbank \mathbf{I} (vom zu Q passenden Schema)

Aufgabe: Berechne $\llbracket Q \rrbracket(\mathbf{I})$

Variante:

NICHT-LEERHEITS-PROBLEM FÜR \mathcal{A} :

Eingabe: Anfrage $Q \in \mathcal{A}$, Datenbank \mathbf{I} (vom zu Q passenden Schema)

Aufgabe: Ist $\llbracket Q \rrbracket(\mathbf{I}) \neq \emptyset$?

Wichtige Fragestellung:

Welche Ressourcen (etwa Zeit, Platz) sind nötig, um diese Probleme zu lösen?

Folie 33

Datenkomplexität und Kombinierte Komplexität

Die Komplexität der Anfrageauswertung kann unter zwei Blickwinkeln betrachtet werden:

1. Anfrage und Datenbank sind Eingabe \rightsquigarrow *Kombinierte Komplexität*
gemessen in n und k , wobei $n = \|\mathbf{I}\|$ und $k = \|Q\|$
2. Anfrage fest, Datenbank ist Eingabe: \rightsquigarrow *Datenkomplexität*
gemessen nur in $n = \|\mathbf{I}\|$

Rechtfertigung für „Datenkomplexität“:

i.d.R. ist die Anfrage kurz, die Datenbank aber sehr groß.

Hierbei sind $\|\mathbf{I}\|$ und $\|Q\|$ geeignete Maße für die Größe von Datenbanken \mathbf{I} und die Länge von Anfragen Q .

Typische Form von Ergebnissen, die im Laufe der Vorlesung bewiesen werden:

- Die Datenkomplexität des Auswertungsproblems der Relationalen Algebra ist in LOGSPACE.
- Die kombinierte Komplexität des Auswertungsproblems der Relationalen Algebra ist PSPACE-vollständig.

Kapitel 3

Konjunktive Anfragen

3.1 Deskriptiver Ansatz: regelbasiert, graphisch und logikbasiert

Regelbasierte Konjunktive Anfragen — Informell

Folie 34

Beispiel-Anfrage:

- (4) Welche Kinos (Name & Adresse) spielen einen Film mit “Matt Damon”?

Andere Formulierung:

Wenn es in *Filme* ein Tupel $(x_{Titel}, x_{Regie}, \text{“Matt Damon”})$ und
in *Programm* ein Tupel $(x_{Kino}, x_{Titel}, x_{Zeit})$ und
in *Kinos* ein Tupel $(x_{Kino}, x_{Adr}, x_{St}, x_{Tel})$ gibt,
dann nimm das Tupel (x_{Kino}, x_{Adr}) in die Antwort auf

Als *regelbasierte konjunktive Anfrage*:

$$\begin{aligned} Ans(x_{Kino}, x_{Adr}) \leftarrow & Filme(x_{Titel}, x_{Regie}, \text{“Matt Damon”}), \\ & Programm(x_{Kino}, x_{Titel}, x_{Zeit}), \\ & Kinos(x_{Kino}, x_{Adr}, x_{St}, x_{Tel}) \end{aligned}$$

Folie 35

Regelbasierte Konjunktive Anfragen — Präzise

Definition 3.1.

- **var** sei eine abzählbar unendliche Menge von *Variablen*(*symbolen*), die disjunkt zu den Mengen **att**, **dom**, **rel** ist.
Einzelne Variablen bezeichnen wir i.d.R. mit x, y, x_1, x_2, \dots
- Ein *Term* ist ein Element aus $\mathbf{var} \cup \mathbf{dom}$.
- Ein *freies Tupel* der Stelligkeit k ist ein Element aus $(\mathbf{var} \cup \mathbf{dom})^k$.

Definition 3.2.

Sei \mathbf{S} ein Datenbankschema.
Eine *regelbasierte konjunktive Anfrage* über \mathbf{S} ist ein Ausdruck der Form

$$Ans(u) \leftarrow R_1(u_1), \dots, R_\ell(u_\ell)$$

wobei $\ell \geq 0$, $R_1, \dots, R_\ell \in \mathbf{S}$, $Ans \in \mathbf{rel} \setminus \mathbf{S}$ und u, u_1, \dots, u_ℓ freie Tupel der Stelligkeiten $\text{ar}(Ans), \text{ar}(R_1), \dots, \text{ar}(R_\ell)$, so dass jede Variable, die in u vorkommt, auch in mindestens einem der Tupel u_1, \dots, u_ℓ vorkommt.

Folie 36

Semantik regelbasierter konjunktiver Anfragen

Sei Q eine regelbasierte konjunktive Anfrage (über einem DB-Schema \mathbf{S}) der Form

$$Ans(u) \leftarrow R_1(u_1), \dots, R_\ell(u_\ell)$$

- Mit $\text{var}(Q)$ bezeichnen wir die Menge aller Variablen, die in Q vorkommen.
- Eine *Belegung* für Q ist eine Abbildung $\beta : \text{var}(Q) \rightarrow \mathbf{dom}$.
Wir setzen β auf natürliche Weise fort zu einer Abbildung von $\text{var}(Q) \cup \mathbf{dom}$ nach \mathbf{dom} , so dass $\beta|_{\mathbf{dom}} = \text{id}$. Für ein freies Tupel $u = (e_1, \dots, e_k)$ setzen wir $\beta(u) := (\beta(e_1), \dots, \beta(e_k))$.
- Der Anfrage Q ordnen wir die folgende Anfragefunktion $\llbracket Q \rrbracket$ zu:

$$\llbracket Q \rrbracket(\mathbf{I}) := \left\{ \begin{array}{l} \beta \text{ ist eine Belegung für } Q, \text{ so} \\ \beta(u) : \text{dass } \beta(u_i) \in \mathbf{I}(R_i), \text{ für alle } i \in \\ \{1, \dots, \ell\} \end{array} \right\}$$

für alle Datenbanken $\mathbf{I} \in \text{inst}(\mathbf{S})$.

Folie 37

Beispiele

- Die Anfrage (6) Welche (je 2) Regisseure haben schon in Filmen gespielt, in denen der jeweils andere Regie geführt hat?

lässt sich durch folgende regelbasierte konjunktive Anfrage ausdrücken:

$$\text{Antworten}(x, y) \leftarrow \text{Filme}(z_1, x, y), \text{Filme}(z_2, y, x)$$

- Die Anfrage (5) Läuft zur Zeit ein Film von “James Cameron”?

lässt sich durch folgende regelbasierte konjunktive Anfrage ausdrücken:

$$\text{Ans}() \leftarrow \text{Filme}(x, \text{“James Cameron”}, y), \text{Programm}(z, x, w)$$

Ans ist hier also ein Relationsname der Stelligkeit 0.

Erinnern Sie sich an unsere Konvention, dass die Ausgabe „ \emptyset “ der Antwort „nein“ entspricht, und die Ausgabe der Menge $\{()\}$, die aus dem „Tupel der Stelligkeit 0“ besteht, der Antwort „ja“ entspricht.

Folie 38

Noch ein Beispiel

Betrachte die Datenbank $\mathbf{I} := \{\mathbf{I}(R), \mathbf{I}(S)\}$ mit

$\mathbf{I}(R) := \{(a, a), (a, b), (b, c), (c, b)\}$ und $\mathbf{I}(S) := \{(d, a, b), (a, c, e), (b, a, c)\}$.

- Die Anfrage $Q_1 :=$

$$\text{Ans}_1(x_1, x_2, x_3) \leftarrow R(x_1, y), S(y, x_2, x_3)$$

liefert auf \mathbf{I} das Ergebnis $\llbracket Q_1 \rrbracket(\mathbf{I}) = \{(a, c, e), (a, a, c), (c, a, c)\}$.

- Die Anfrage $Q_2 :=$

$$\text{Ans}_2(x, y) \leftarrow R(x, z_1), S(z_1, a, z_2), R(y, z_2)$$

liefert auf \mathbf{I} das Ergebnis $\llbracket Q_2 \rrbracket(\mathbf{I}) = \{(a, b), (c, b)\}$.

Folie 39

Bezeichnungen

Oft sagen wir kurz *Regel*, um eine regelbasierte konjunktive Anfrage zu bezeichnen.

Sei Q eine Regel der Form $Ans(u) \leftarrow R_1(u_1), \dots, R_\ell(u_\ell)$

- $Ans(u)$ wird als *Kopf* der Regel bezeichnet.
- $R_1(u_1), \dots, R_\ell(u_\ell)$ wird als *Rumpf* der Regel bezeichnet.
- Die Relationsnamen aus \mathbf{S} werden *extensionale Datenbankprädikate* (kurz: *edb-Prädikate*) genannt.
Wir schreiben $edb(Q)$, um die Menge aller edb-Prädikate zu bezeichnen, die in Q vorkommen.
- Der Relationsname, der im Kopf von Q vorkommt, wird als *intensionales Datenbankprädikat* (kurz: *idb-Prädikat*) bezeichnet.
- Mit $adom(Q)$ bezeichnen wir die Menge aller *Konstanten* (also Elemente aus \mathbf{dom}), die in Q vorkommen.

„ $adom$ “ steht für „*aktive Domäne*“ bzw. „*active domain*“.

Folie 40

Der „Active Domain“ einer Datenbank

Definition 3.3.

Sei \mathbf{S} ein Datenbankschema und sei \mathbf{I} eine Datenbank vom Schema \mathbf{S} .

Der *Active Domain* von \mathbf{I} , kurz: $adom(\mathbf{I})$, ist die Menge aller Elemente aus \mathbf{dom} , die in \mathbf{I} vorkommen. D.h.:

$$adom(\mathbf{I}) = \bigcup_{R \in \mathbf{S}} adom(\mathbf{I}(R))$$

wobei für jedes R aus \mathbf{S} gilt: $adom(\mathbf{I}(R))$ ist die kleinste Teilmenge von \mathbf{dom} , so dass jedes Tupel $t \in \mathbf{I}(R)$ eine Funktion von $sorte(R)$ nach $adom(\mathbf{I}(R))$ ist.

Ist Q eine Anfrage und \mathbf{I} eine Datenbank, so setzen wir

$$adom(Q, \mathbf{I}) := adom(Q) \cup adom(\mathbf{I})$$

Proposition 3.4.

Für jede regelbasierte konjunktive Anfrage Q und jede Datenbank \mathbf{I} (vom passenden DB-Schema) gilt: $\text{adom}(\llbracket Q \rrbracket(\mathbf{I})) \subseteq \text{adom}(Q, \mathbf{I})$.

Beweis. Sei Q von der Form

$$\text{Ans}(u) \leftarrow R_1(u_1), \dots, R_\ell(u_\ell).$$

Sei $r := \text{ar}(\text{Ans})$ und $u = (x_1, \dots, x_r)$. Insbesondere ist jeder Eintrag x_j von u ein Element aus $\mathbf{var} \cup \mathbf{dom}$.

Sei $t := (t_1, \dots, t_r)$ ein beliebiges Element aus $\llbracket Q \rrbracket(\mathbf{I})$. Wir müssen zeigen, dass jeder Eintrag t_j von t ein Element in $\text{adom}(Q, \mathbf{I})$ ist.

Wegen $t \in \llbracket Q \rrbracket(\mathbf{I})$ gibt es gemäß der Definition der Semantik regelbasierter konjunktiver Anfragen eine Belegung β für Q , so dass gilt:

- (a) $t = \beta(u)$ und
- (b) $\beta(u_i) \in \mathbf{I}(R_i)$, für jedes $i \in [\ell] := \{1, \dots, \ell\}$.

Wegen (a) gilt also für jedes $j \in [r]$, dass $t_j = \beta(x_j)$.

Fall 1: $x_j \in \mathbf{dom}$.

Dann gilt: $t_j = \beta(x_j) = x_j \in \text{adom}(Q) \subseteq \text{adom}(Q, \mathbf{I})$.

Fall 2: $x_j \in \mathbf{var}$.

Gemäß der Definition der Syntax regelbasierter konjunktiver Anfragen (Definition 3.2) gibt es ein $i \in [\ell]$, so dass x_j im freien Tupel u_i vorkommt.

Dann kommt also $\beta(x_j)$ im Tupel $\beta(u_i)$ vor. Gemäß (b) wissen wir, dass $\beta(u_i) \in \mathbf{I}(R_i)$ ist. Somit ist $\beta(x_j) \in \text{adom}(\mathbf{I}) \subseteq \text{adom}(Q, \mathbf{I})$. □

Folie 41

Monotonie und Erfüllbarkeit

Sind \mathbf{I} und \mathbf{J} zwei Datenbanken vom gleichen Schema \mathbf{S} , so sagen wir „ \mathbf{J} ist eine Erweiterung von \mathbf{I} “ und schreiben kurz „ $\mathbf{J} \supseteq \mathbf{I}$ “ (bzw. „ $\mathbf{I} \subseteq \mathbf{J}$ “), falls für alle $R \in \mathbf{S}$ gilt: $\mathbf{I}(R) \subseteq \mathbf{J}(R)$ (d.h. jedes Tupel, das in einer Relation von \mathbf{I} vorkommt, kommt auch in der entsprechenden Relation von \mathbf{J} vor).

Definition 3.5.

Sei \mathbf{S} ein DB-Schema und sei q eine Anfragefunktion über \mathbf{S} .

- (a) q heißt *monoton*, falls für alle Datenbanken \mathbf{I} und \mathbf{J} über \mathbf{S} gilt:
Falls $\mathbf{I} \subseteq \mathbf{J}$, so ist $q(\mathbf{I}) \subseteq q(\mathbf{J})$.

(b) q heißt *erfüllbar*, falls es eine Datenbank \mathbf{I} gibt mit $q(\mathbf{I}) \neq \emptyset$.

Satz 3.6. *Jede regelbasierte konjunktive Anfrage ist monoton und erfüllbar.*

Beweis. Sei \mathbf{S} ein DB-Schema und sei Q eine Anfrage über \mathbf{S} der Form

$$\text{Ans}(u) \leftarrow R_1(u_1), \dots, R_\ell(u_\ell).$$

Nachweis der Monotonie von Q :

Seien \mathbf{I} und \mathbf{J} Datenbanken mit $\mathbf{I} \subseteq \mathbf{J}$.

Wir müssen zeigen, dass $\llbracket Q \rrbracket(\mathbf{I}) \subseteq \llbracket Q \rrbracket(\mathbf{J})$ ist.

Sei also $t \in \llbracket Q \rrbracket(\mathbf{I})$ beliebig gewählt. Zu zeigen: $t \in \llbracket Q \rrbracket(\mathbf{J})$.

Wegen $t \in \llbracket Q \rrbracket(\mathbf{I})$ gibt es eine Belegung β für Q , so dass $t = \beta(u)$ ist und für alle $i \in [\ell]$ gilt $\beta(u_i) \in \mathbf{I}(R_i)$.

Wegen $\mathbf{I} \subseteq \mathbf{J}$ gilt $\mathbf{I}(R_i) \subseteq \mathbf{J}(R_i)$ für jedes $i \in [\ell]$.

Somit gilt für jedes $i \in [\ell]$, dass $\beta(u_i) \in \mathbf{I}(R_i) \subseteq \mathbf{J}(R_i)$. Daher ist β eine Belegung, die gemäß der Definition der Semantik regelbasierter konjunktiver Anfragen bezeugt, dass $\beta(u) \in \llbracket Q \rrbracket(\mathbf{J})$ ist. Also ist $t \in \llbracket Q \rrbracket(\mathbf{J})$.

Nachweis der Erfüllbarkeit von Q :

Wir müssen eine Datenbank \mathbf{I} vom Schema \mathbf{S} finden, für die $\llbracket Q \rrbracket(\mathbf{I}) \neq \emptyset$ ist.

Dazu wählen wir eine Konstante $a \in \mathbf{dom} \setminus \text{adom}(Q)$ (die gibt es, da $\text{adom}(Q)$ endlich und \mathbf{dom} unendlich ist).

Für jedes $R \in \mathbf{S}$ setze

$$\mathbf{I}(R) := \left(\text{adom}(Q) \cup \{a\} \right)^{\text{ar}(R)}.$$

Sei β die Belegung für Q , die jede Variable aus Q auf das Element a abbildet.

Dann ist $\beta(u) \in \llbracket Q \rrbracket(\mathbf{I})$, denn für jedes $i \in [\ell]$ ist

$$\beta(u_i) \in \underbrace{\left(\text{adom}(Q) \cup \{a\} \right)^{\text{ar}(R_i)}}_{= \mathbf{I}(R_i)}.$$

Insbesondere ist $\llbracket Q \rrbracket(\mathbf{I}) \neq \emptyset$, d.h. Q ist erfüllbar. □

Anwendung von Satz 3.6

Satz 3.6 liefert ein einfaches Kriterium, um zu zeigen, dass bestimmte Anfragefunktionen nicht durch eine regelbasierte konjunktive Anfrage beschrieben werden können:

Wenn eine Anfragefunktion q *nicht monoton* ist, dann kann sie auch *nicht durch eine regelbasierte konjunktive Aussage beschrieben* werden.

Vorsicht: Dies heißt nicht, dass jede monotone Anfragefunktion durch eine regelbasierte konjunktive Anfrage beschrieben werden kann!

Beispiel: Die Anfrage

(15) Welche Filme laufen nur zu 1 Uhrzeit?

ist nicht monoton, kann also nicht durch eine regelbasierte konjunktive Anfrage beschrieben werden.

Dass die Anfrage (15) nicht monoton ist, sieht man z.B. dadurch, dass man die Datenbank \mathbf{I} betrachtet, die aus unserer Beispiel-Datenbank \mathbf{I}_{KINO} entsteht, indem das Tupel (Babylon, Casablanca, 17:30) aus der *Programm*-Relation gelöscht wird. Offensichtlicherweise ist $\mathbf{I} \subseteq \mathbf{I}_{\text{KINO}}$. Aber das Ergebnis der Anfrage (15) ist bei Auswertung über der Datenbank \mathbf{I}_{KINO} leer, während es bei Auswertung über der Datenbank \mathbf{I} den Film “Casablanca” liefert (der gemäß der Datenbank \mathbf{I} nur um 18:00 Uhr läuft).

Folie 43

„Graphische“ Variante: Tableau-Anfragen

Beispiel-Anfrage:

(4) Welche Kinos (Name & Adresse) spielen einen Film mit “Matt Damon”?

Darstellung als Tableau T (ähnlich zu „Query by Example“ (QBE) von IBM):

<i>Filme</i>	<table style="border-collapse: collapse; border: none;"> <tr> <td style="padding: 0 10px;">Titel</td> <td style="padding: 0 10px;">Regie</td> <td style="padding: 0 10px;">Schauspieler</td> </tr> <tr> <td style="padding: 0 10px;">x_{Titel}</td> <td style="padding: 0 10px;">x_{Regie}</td> <td style="padding: 0 10px;">“Matt Damon”</td> </tr> </table>	Titel	Regie	Schauspieler	x_{Titel}	x_{Regie}	“Matt Damon”		
Titel	Regie	Schauspieler							
x_{Titel}	x_{Regie}	“Matt Damon”							
<i>Programm</i>	<table style="border-collapse: collapse; border: none;"> <tr> <td style="padding: 0 10px;">Kino</td> <td style="padding: 0 10px;">Titel</td> <td style="padding: 0 10px;">Zeit</td> </tr> <tr> <td style="padding: 0 10px;">x_{Kino}</td> <td style="padding: 0 10px;">x_{Titel}</td> <td style="padding: 0 10px;">x_{Zeit}</td> </tr> </table>	Kino	Titel	Zeit	x_{Kino}	x_{Titel}	x_{Zeit}		
Kino	Titel	Zeit							
x_{Kino}	x_{Titel}	x_{Zeit}							
<i>Kinos</i>	<table style="border-collapse: collapse; border: none;"> <tr> <td style="padding: 0 10px;">Kino</td> <td style="padding: 0 10px;">Adresse</td> <td style="padding: 0 10px;">Stadtteil</td> <td style="padding: 0 10px;">Telefon</td> </tr> <tr> <td style="padding: 0 10px;">x_{Kino}</td> <td style="padding: 0 10px;">x_{Adr}</td> <td style="padding: 0 10px;">x_{St}</td> <td style="padding: 0 10px;">x_{Tel}</td> </tr> </table>	Kino	Adresse	Stadtteil	Telefon	x_{Kino}	x_{Adr}	x_{St}	x_{Tel}
Kino	Adresse	Stadtteil	Telefon						
x_{Kino}	x_{Adr}	x_{St}	x_{Tel}						

Zugehörige Tableau-Anfrage: $(\mathbb{T}, (x_{\text{Kino}}, x_{\text{Adr}}))$

Folie 44

Tableaus — Präzise

Definition 3.7. Sei \mathbf{S} ein Datenbankschema und R ein Relationsschema.

- Ein *Tableau über R* (auch: Einzel-Tableau) ist eine endliche Menge von freien Tupeln (also Tupeln über $\mathbf{dom} \cup \mathbf{var}$) der Stelligkeit $\text{ar}(R)$.
(D.h. ein Tableau über R ist eine „Relation vom Schema R , die als Einträge nicht nur Elemente aus \mathbf{dom} , sondern auch Variablen aus \mathbf{var} haben kann“.)
- Ein *Tableau \mathbb{T} über \mathbf{S}* ist eine Abbildung, die jedem $R \in \mathbf{S}$ ein Tableau über R zuordnet.
(D.h. ein Tableau über \mathbf{S} ist eine „Datenbank vom Schema \mathbf{S} , die als Einträge auch Variablen enthalten kann“.)
- Eine *Tableau-Anfrage über \mathbf{S}* (bzw. R) ist von der Form (\mathbb{T}, u) , wobei \mathbb{T} ein Tableau über \mathbf{S} (bzw. R) und u ein freies Tupel ist, so dass jede Variable, die in u vorkommt, auch in \mathbb{T} vorkommt.
 u heißt *Zusammenfassung* der Anfrage (\mathbb{T}, u) .

Folie 45

Semantik von Tableau-Anfragen

Sei $Q = (\mathbb{T}, u)$ eine Tableau-Anfrage.

- $\text{var}(Q)$ bezeichnet die Menge aller Variablen, die in u oder \mathbb{T} vorkommen.
 $\text{adom}(Q)$ bezeichnet die Menge aller Konstanten, die in u oder \mathbb{T} vorkommen.
- Eine *Belegung für Q* ist eine Abbildung $\beta : \text{var}(Q) \rightarrow \mathbf{dom}$.
Wir schreiben $\beta(\mathbb{T})$, um die Datenbank zu bezeichnen, die aus \mathbb{T} entsteht, indem man jedes Variablensymbol x in \mathbb{T} durch die Konstante $\beta(x)$ ersetzt.
- Sei \mathbf{I} eine Datenbank vom Schema \mathbf{S} .
Eine Belegung β für Q heißt *Einbettung von \mathbb{T} in \mathbf{I}* , falls $\beta(\mathbb{T}) \subseteq \mathbf{I}$ gilt (d.h. die Datenbank \mathbf{I} ist eine Erweiterung der Datenbank $\beta(\mathbb{T})$).

- Der Tableau-Anfrage Q ordnen wir die folgende Anfragefunktion $\llbracket Q \rrbracket$ zu:

$$\llbracket Q \rrbracket(\mathbf{I}) := \{ \beta(u) : \beta \text{ ist eine Einbettung von } \mathbf{T} \text{ in } \mathbf{I} \}$$

für alle Datenbanken $\mathbf{I} \in \text{inst}(\mathbf{S})$.

Folie 46

Logikbasierte Variante: Konjunktiver Kalkül

Beispiel-Anfrage:

- (4) Welche Kinos (Name & Adresse) spielen einen Film mit “Matt Damon”?

Formulierung als Anfrage des konjunktiven Kalküls:

$$\left\{ \begin{array}{l} (x_{Kino}, x_{Adr}) : \exists x_{Titel} \exists x_{Regie} \exists x_{Zeit} \exists x_{St} \exists x_{Tel} (\\ \quad \text{Filme}(x_{Titel}, x_{Regie}, \text{“Matt Damon”}) \wedge \\ \quad \text{Programm}(x_{Kino}, x_{Titel}, x_{Zeit}) \wedge \\ \quad \text{Kinოს}(x_{Kino}, x_{Adr}, x_{St}, x_{Tel}) \end{array} \right\}$$

Auf der rechten Seite des „:“

werden Varianten von Formeln der *Logik erster Stufe* verwendet.

Hier: eingeschränkte Variante, in der es nur \exists -Quantoren und Konjunktionen (\wedge) gibt.

Folie 47

Konjunktiver Kalkül (CQ) — Präzise

Definition 3.8. Sei \mathbf{S} ein Datenbankschema.

Die Menge $\text{CQ}[\mathbf{S}]$ aller Formeln des *konjunktiven Kalküls über \mathbf{S}* ist

induktiv wie folgt definiert: (*CQ* steht für „conjunctive queries“)

- (A) $R(v_1, \dots, v_r)$ gehört zu $\text{CQ}[\mathbf{S}]$,
für alle $R \in \mathbf{S}$, $r := \text{ar}(R)$ und $v_1, \dots, v_r \in \mathbf{var} \cup \mathbf{dom}$.
- (K) $(\varphi \wedge \psi)$ gehört zu $\text{CQ}[\mathbf{S}]$,
für alle $\varphi \in \text{CQ}[\mathbf{S}]$ und $\psi \in \text{CQ}[\mathbf{S}]$.
- (E) $\exists x \varphi$ gehört zu $\text{CQ}[\mathbf{S}]$,
für alle $x \in \mathbf{var}$ und $\varphi \in \text{CQ}[\mathbf{S}]$.

Insbesondere: Jede Formel in $\text{CQ}[\mathbf{S}]$ ist eine Formel der *Logik erster Stufe* über der Signatur $\mathbf{S} \cup \mathbf{dom}$ (wobei jedes Element aus \mathbf{dom} als „Konstanten-Symbol“ aufgefasst wird, das stets „mit sich selbst“ interpretiert wird).

Folie 48

Semantik von $\text{CQ}[\mathbf{S}]$

Sei φ eine $\text{CQ}[\mathbf{S}]$ -Formel.

- $\text{adom}(\varphi)$ bezeichnet die Menge aller Konstanten (also Elemente aus \mathbf{dom}), die in φ vorkommen.
 $\text{var}(\varphi)$ bezeichnet die Menge aller Variablen (also Elemente aus \mathbf{var}), die in φ vorkommen.
- $\text{frei}(\varphi)$ bezeichnet die Menge aller Variablen, die *frei* in φ vorkommen.
D.h.: $\text{frei}(R(v_1, \dots, v_r)) = \{v_1, \dots, v_r\} \cap \mathbf{var}$;
 $\text{frei}(\varphi \wedge \psi) = \text{frei}(\varphi) \cup \text{frei}(\psi)$;
 $\text{frei}(\exists x \varphi) = \text{frei}(\varphi) \setminus \{x\}$.
- Eine *Belegung* für φ ist eine Abbildung $\beta : \text{frei}(\varphi) \rightarrow \mathbf{dom}$.
- Einer *Datenbank* \mathbf{I} vom Schema \mathbf{S} ordnen wir die *logische Struktur*

$$\mathcal{A}_{\mathbf{I}} := \left(\mathbf{dom}, (\mathbf{I}(R))_{R \in \mathbf{S}}, (c)_{c \in \mathbf{dom}} \right)$$

zu. *Insbesondere ist $\mathcal{A}_{\mathbf{I}}$ eine σ -Struktur über der Signatur $\sigma := \mathbf{S} \cup \mathbf{dom}$.*

- Ist \mathbf{I} eine Datenbank vom Schema \mathbf{S} und β eine Belegung für φ , so sagen wir „ \mathbf{I} erfüllt φ unter β “ und schreiben $\mathbf{I} \models \varphi[\beta]$ bzw. $(\mathbf{I}, \beta) \models \varphi$, um auszudrücken, dass $(\mathcal{A}_{\mathbf{I}}, \beta) \models \varphi$ gilt.

Bemerkung. Wir verwenden hier Notationen, die in der Veranstaltung *Logik in der Informatik* eingeführt wurden. Der Vollständigkeit halber geben wir hier die Definition rekursiv für alle $\text{CQ}[\mathbf{S}]$ -Formeln φ an:

- Falls φ von der Form $R(v_1, \dots, v_r)$ ist, für $R \in \mathbf{S}$, $r = \text{ar}(R)$ und $v_1, \dots, v_r \in \text{var} \cup \mathbf{dom}$, so gilt

$$(\mathcal{A}_{\mathbf{I}}, \beta) \models \varphi \quad :\iff \quad (\beta(v_1), \dots, \beta(v_r)) \in \mathbf{I}(R)$$

- Falls φ von der Form $(\psi_1 \wedge \psi_2)$ ist, für $\psi_1 \in \text{CQ}[\mathbf{S}]$ und $\psi_2 \in \text{CQ}[\mathbf{S}]$, so gilt

$$(\mathcal{A}_{\mathbf{I}}, \beta) \models \varphi \quad :\iff \quad ((\mathcal{A}_{\mathbf{I}}, \beta) \models \psi_1 \quad \text{und} \quad (\mathcal{A}_{\mathbf{I}}, \beta) \models \psi_2)$$

- Falls φ von der Form $\exists x \psi$ ist, für $x \in \text{var}$ und $\psi \in \text{CQ}[\mathbf{S}]$, so gilt

$$(\mathcal{A}_{\mathbf{I}}, \beta) \models \varphi \quad :\iff \quad \text{es gibt ein } a \in \mathbf{dom}, \text{ so dass für die Belegung } \beta_x^a : \text{frei}(\varphi) \rightarrow \mathbf{dom} \text{ mit } \beta_x^a(x) := a \text{ und } \beta_x^a(y) := \beta(y) \text{ für alle } y \neq x \text{ gilt: } (\mathcal{A}_{\mathbf{I}}, \beta_x^a) \models \psi$$

Folie 49

Notation

- Mit CQ bezeichnen wir die Klasse aller $\text{CQ}[\mathbf{S}]$ -Formeln für alle Datenbankschemata \mathbf{S} .
- Manchmal schreiben wir $\{x_1/a_1, \dots, x_r/a_r\}$ um die *Belegung* $\beta : \{x_1, \dots, x_r\} \rightarrow \mathbf{dom}$ zu bezeichnen mit $\beta(x_i) = a_i$, f.a. $i \in \{1, \dots, r\}$.
- Für eine CQ -Formel φ schreiben wir oft $\varphi(x_1, \dots, x_r)$, um anzudeuten, dass $\text{frei}(\varphi) \subseteq \{x_1, \dots, x_r\}$.
- Ist $a_1, \dots, a_r \in \mathbf{dom}$, so schreiben wir vereinfachend $\mathbf{I} \models \varphi[a_1, \dots, a_r]$ an Stelle von $\mathbf{I} \models \varphi[\{x_1/a_1, \dots, x_r/a_r\}]$.
- Ist $y \in \mathbf{dom} \cup \text{var}$, so bezeichnet $\varphi(x_1/y, x_2, \dots, x_r)$ die Formel, die aus φ entsteht, indem jedes Vorkommen der *Variablen* x_1 durch y ersetzt wird.
- Beim Schreiben von Formeln lassen wir Klammern „(, „)“ oft weg und schreiben „ $\exists x_1, \dots, x_n$ “ als Abkürzung für „ $\exists x_1 \exists x_2 \dots \exists x_n$ “.
- Zwei $\text{CQ}[\mathbf{S}]$ -Formeln φ und ψ heißen *äquivalent*, falls $\text{frei}(\varphi) = \text{frei}(\psi)$ und für jede Datenbank $\mathbf{I} \in \text{inst}(\mathbf{S})$ und jede Belegung β für φ (und ψ) gilt: $\mathbf{I} \models \varphi[\beta] \iff \mathbf{I} \models \psi[\beta]$.

Folie 50

Konjunktiver Kalkül: Syntax und Semantik

Definition 3.9. Sei \mathbf{S} ein Datenbankschema.

Eine *Anfrage des konjunktiven Kalküls* ist von der Form

$$\{(e_1, \dots, e_r) : \varphi\}$$

wobei $\varphi \in \text{CQ}[\mathbf{S}]$, $r \geq 0$ und (e_1, \dots, e_r) ein freies Tupel ist, so dass $\text{frei}(\varphi) = \{e_1, \dots, e_r\} \cap \mathbf{var}$.

Semantik:

Einer Anfrage Q der Form $\{(e_1, \dots, e_r) : \varphi\}$ ordnen wir die folgende Anfragefunktion $\llbracket Q \rrbracket$ zu:

$$\llbracket Q \rrbracket(\mathbf{I}) := \left\{ \beta((e_1, \dots, e_r)) : \begin{array}{l} \beta \text{ ist eine Belegung für } \varphi \\ \text{mit } \mathbf{I} \models \varphi[\beta] \end{array} \right\}$$

für alle Datenbanken $\mathbf{I} \in \text{inst}(\mathbf{S})$.

Folie 51

Wertebereich von Anfragen des konjunktiven Kalküls

Für eine Anfrage Q der Form $\{(e_1, \dots, e_r) : \varphi\}$ setzen wir

$$\text{adom}(Q) := \text{adom}(\varphi) \cup (\{e_1, \dots, e_r\} \cap \mathbf{dom}).$$

Ist Q eine Anfrage und \mathbf{I} eine Datenbank, so setzen wir – wie üblich –

$$\text{adom}(Q, \mathbf{I}) := \text{adom}(Q) \cup \text{adom}(\mathbf{I})$$

Analog zu Proposition 3.4 gilt auch für Anfragen des konjunktiven Kalküls:

Proposition 3.10.

Für jede Anfrage Q des konjunktiven Kalküls und jede Datenbank \mathbf{I} (vom passenden DB-Schema) gilt: $\text{adom}(\llbracket Q \rrbracket(\mathbf{I})) \subseteq \text{adom}(Q, \mathbf{I})$.

Beweis: Induktion über den Formelaufbau. Details: Übung.

Folie 52

Beispiel

Die Anfrage

Gibt es einen Schauspieler, der sowohl in einem Film von “Paul Anderson” als auch in einem Film von “Ridley Scott” mitgespielt hat?

wird durch die folgende Anfrage des konjunktiven Kalküls beschrieben:

$$\left\{ () : \exists x_{\text{Schauspieler}} \left(\exists x_{\text{Titel1}} \text{Filme}(x_{\text{Titel1}}, \text{“Paul Anderson”}, x_{\text{Schauspieler}}) \wedge \exists x_{\text{Titel2}} \text{Filme}(x_{\text{Titel2}}, \text{“Ridley Scott”}, x_{\text{Schauspieler}}) \right) \right\}$$

und durch die dazu äquivalente Anfrage

$$\left\{ () : \exists x_{\text{Schauspieler}} \exists x_{\text{Titel1}} \exists x_{\text{Titel2}} \left(\text{Filme}(x_{\text{Titel1}}, \text{“Paul Anderson”}, x_{\text{Schauspieler}}) \wedge \text{Filme}(x_{\text{Titel2}}, \text{“Ridley Scott”}, x_{\text{Schauspieler}}) \right) \right\}$$

Folie 53

Eine Normalform für CQ

Definition 3.11.

Eine CQ[S]-Formel $\varphi(x_1, \dots, x_r)$ ist in Normalform, falls sie von der Form

$$\exists x_{r+1} \dots \exists x_{r+s} \left(R_1(u_1) \wedge \dots \wedge R_\ell(u_\ell) \right)$$

ist, wobei gilt: $r, s, \ell \geq 0$, $R_1, \dots, R_\ell \in \mathbf{S}$, u_1, \dots, u_ℓ sind freie Tupel über $\{x_1, \dots, x_{r+s}\} \cup \mathbf{dom}$, $\text{frei}(\varphi) = \{x_1, \dots, x_r\}$ und x_1, \dots, x_{r+s} sind paarweise verschiedene Elemente aus \mathbf{var} .

Eine Anfrage Q des konjunktiven Kalküls ist in Normalform, falls sie von der Form $\{(e_1, \dots, e_r) : \varphi\}$ ist, wobei φ eine CQ-Formel in Normalform ist.

Lemma 3.12.

Jede CQ-Formel ist äquivalent zu einer CQ-Formel in Normalform.

Und es gibt einen Polynomialzeit-Algorithmus, der bei Eingabe einer CQ-Formel eine äquivalente CQ-Formel in Normalform konstruiert.

Beweis: Übung.

Folie 54

Beispiel

Die CQ-Formel

$$\begin{aligned} & \exists z \exists z' \text{ Programm}(z, x, z') \wedge \\ & \exists z \text{ Filme}(x, y, z) \wedge \\ & \exists z \text{ Filme}(z, y, \text{“George Clooney”}) \wedge \\ & \exists z \text{ Filme}(z, \text{“George Clooney”}, y) \end{aligned}$$

ist nicht in Normalform, aber äquivalent zur Normalform-Formel

$$\begin{aligned} \exists z_1 \cdots \exists z_5 (& \text{ Programm}(z_1, x, z_2) \wedge \\ & \text{ Filme}(x, y, z_3) \wedge \\ & \text{ Filme}(z_4, y, \text{“George Clooney”}) \wedge \\ & \text{ Filme}(z_5, \text{“George Clooney”}, y) \end{aligned}$$

wobei z_1, \dots, z_5 paarweise verschiedene Elemente aus **var** sind.

Folie 55

Äquivalenz von Anfragesprachen

Definition 3.13. Seien \mathcal{Q}_1 und \mathcal{Q}_2 zwei Anfragesprachen. Wir schreiben

- $\mathcal{Q}_1 \leq \mathcal{Q}_2$ (bzw. $\mathcal{Q}_2 \geq \mathcal{Q}_1$; „ \mathcal{Q}_2 ist mindestens so ausdrucksstark wie \mathcal{Q}_1 “), falls jede Anfragefunktion, die durch eine Anfrage in \mathcal{Q}_1 ausgedrückt werden kann, auch durch eine Anfrage in \mathcal{Q}_2 ausgedrückt werden kann.
- $\mathcal{Q}_1 \equiv \mathcal{Q}_2$ („ \mathcal{Q}_1 und \mathcal{Q}_2 haben dieselbe Ausdrucksstärke“) falls $\mathcal{Q}_1 \leq \mathcal{Q}_2$ und $\mathcal{Q}_2 \leq \mathcal{Q}_1$
- $\mathcal{Q}_1 < \mathcal{Q}_2$ (bzw. $\mathcal{Q}_2 > \mathcal{Q}_1$; „ \mathcal{Q}_2 ist ausdrucksstärker als \mathcal{Q}_1 “) falls $\mathcal{Q}_1 \leq \mathcal{Q}_2$ und nicht $\mathcal{Q}_2 \leq \mathcal{Q}_1$.

Folie 56

Äquivalenz der bisher eingeführten Anfragesprachen

Lemma 3.14.

Die Klassen der regelbasierten konjunktiven Anfragen, der Tableau-Anfragen und der Anfragen des konjunktiven Kalküls haben dieselbe Ausdrucksstärke.

Es gilt sogar: Jede Anfrage aus einer dieser drei Anfragesprachen kann in polynomieller Zeit in äquivalente Anfragen der beiden anderen Anfragesprachen übersetzt werden.

Beweis.

Regelbasiert \rightsquigarrow *Tableau*:

Gegeben sei eine regelbasierte konjunktive Anfrage Q der Form

$$\text{Ans}(u) \leftarrow R_1(u_1), \dots, R_\ell(u_\ell)$$

Sei $Q' = (\mathsf{T}', u')$ die Tableau-Anfrage mit $u' := u$ und

$$\mathsf{T}'(R) := \{u_i : i \in \{1, \dots, \ell\} \text{ mit } R_i = R\},$$

für jedes $R \in \mathbf{S}$.

Einfaches Nachrechnen liefert für alle Datenbanken $\mathbf{I} \in \text{inst}(\mathbf{S})$, dass $\llbracket Q' \rrbracket(\mathbf{I}) = \llbracket Q \rrbracket(\mathbf{I})$ ist:

Für die Richtung „ \subseteq “ sei t ein beliebiges Tupel aus $\llbracket Q' \rrbracket(\mathbf{I})$. Ziel ist, zu zeigen, dass $t \in \llbracket Q \rrbracket(\mathbf{I})$ ist.

Wegen $t \in \llbracket Q' \rrbracket(\mathbf{I})$ gibt es laut Definition der Semantik von

Tableau-Anfragen eine Einbettung β von Q' in \mathbf{I} , so dass $t = \beta(u')$ ist.

Gemäß Definition des Begriffs einer „Einbettung“ ist β eine Abbildung $\beta : \text{var}(Q') \rightarrow \mathbf{dom}$, für die gilt: $\beta(\mathsf{T}') \subseteq \mathbf{I}$, d.h. für jedes $R \in \mathbf{S}$ ist $\beta(\mathsf{T}'(R)) \subseteq \mathbf{I}(R)$.

Gemäß Konstruktion von Q' ist $\text{var}(Q') = \text{var}(Q)$ und $u' = u$. Wir wollen nun zeigen, dass für jedes Atom $R_i(u_i)$ im Rumpf von Q gilt: $\beta(u_i) \in \mathbf{I}(R_i)$.

Beachte: Gemäß der Definition der Semantik regelbasierter konjunktiver Anfragen bezeugt β dann, dass $\beta(u) \in \llbracket Q \rrbracket(\mathbf{I})$ ist; und wegen

$\beta(u) = \beta(u') = t$ sind wir dann fertig mit dem Beweis der Richtung „ \subseteq “.

Betrachte nun ein beliebiges Atom $R_i(u_i)$ im Rumpf von Q . Für $R := R_i$ gilt gemäß unserer Wahl des Tableaus T' , dass $u_i \in \mathsf{T}'(R)$ ist. Da β eine Einbettung in \mathbf{I} ist, gilt: $\beta(\mathsf{T}'(R)) \subseteq \mathbf{I}(R)$. Insbesondere ist also $\beta(u_i) \in \mathbf{I}(R)$ für $R := R_i$. Dies beendet den Beweis der Richtung „ \subseteq “.

Für die Richtung „ \supseteq “ sei t ein beliebiges Tupel aus $\llbracket Q \rrbracket(\mathbf{I})$. Ziel ist, zu zeigen, dass $t \in \llbracket Q' \rrbracket(\mathbf{I})$ ist.

Wegen $t \in \llbracket Q \rrbracket(\mathbf{I})$ gibt es laut der Definition der Semantik regelbasierter konjunktiver Anfragen eine Belegung $\beta : \text{var}(Q) \rightarrow \mathbf{dom}$, so dass $t = \beta(u)$ ist und $\beta(u_i) \in \mathbf{I}(R_i)$ für jedes $i \in \{1, \dots, \ell\}$ gilt.

Gemäß der Konstruktion von Q' ist $\text{var}(Q') = \text{var}(Q)$ und $u' = u$, also

$\beta(u') = \beta(u) = t$. Um den Beweis zu beenden, genügt es gemäß der

Definition der Semantik von Tableau-Anfragen also nachzuweisen, dass β

eine Einbettung von T' in \mathbf{I} ist. Wir müssen daher nur noch zeigen, dass

$\beta(\mathsf{T}'(R)) \subseteq \mathbf{I}(R)$ für jedes $R \in \mathbf{S}$ gilt. Betrachte dazu ein beliebiges Element

$R \in \mathbf{S}$ und ein beliebiges Element $v \in \mathsf{T}'(R)$. Ziel ist, zu zeigen, dass

$\beta(v) \in \mathbf{I}(R)$ ist.

Gemäß unserer Wahl des Tableaus \mathbb{T}' gibt es für jedes $v \in \mathbb{T}'(R)$ ein $i \in \{1, \dots, \ell\}$, so dass $v = u_i$ und $R = R_i$. Gemäß Voraussetzung wissen wir, dass gilt: $\beta(u_i) \in \mathbf{I}(R_i)$. Somit gilt also $\beta(v) \in \mathbf{I}(R)$. Dies beendet den Beweis der Richtung „ \supseteq “.

Tableau \rightsquigarrow Konjunktiver Kalkül:

Gegeben sei eine Tableau-Anfrage $Q = (\mathbb{T}, u)$.

Sei x_1, \dots, x_k eine Aufzählung aller Variablen aus $\text{var}(Q)$, die *nicht* im Zusammenfassungstupel u vorkommen.

Sei Q' die folgende Anfrage des Konjunktiven Kalküls:

$$\left\{ (u) : \exists x_1 \cdots \exists x_k \bigwedge_{R \in \mathbf{S}} \bigwedge_{u_R \in \mathbb{T}(R)} R(u_R) \right\}$$

Einfaches Nachrechnen liefert für alle Datenbanken $\mathbf{I} \in \text{inst}(\mathbf{S})$, dass $\llbracket Q' \rrbracket(\mathbf{I}) = \llbracket Q \rrbracket(\mathbf{I})$ ist (Details: Übung).

Konjunktiver Kalkül \rightsquigarrow Regelbasiert:

Gegeben sei eine Anfrage $Q = \{ (u) : \varphi \}$ des Konjunktiven Kalküls.

Wir nutzen Lemma 3.12, um φ in eine äquivalente CQ-Formel φ' in Normalform zu transformieren. Sei φ' von der Form

$$\exists z_1 \cdots \exists z_s (R_1(u_1) \wedge \cdots \wedge R_\ell(u_\ell))$$

Sei Q' die regelbasierte konjunktive Anfrage

$$\text{Ans}(u) \leftarrow R_1(u_1), \dots, R_\ell(u_\ell)$$

Einfaches Nachrechnen liefert für alle Datenbanken $\mathbf{I} \in \text{inst}(\mathbf{S})$, dass $\llbracket Q' \rrbracket(\mathbf{I}) = \llbracket Q \rrbracket(\mathbf{I})$ (Details: Übung). □

Beispiel zur Übersetzung von Anfragen

Betrachte die Anfrage

$$\begin{aligned} \text{Ans}(x, y) \leftarrow & \text{Programm}(z_1, x, z_2), \\ & \text{Filme}(x, y, z_3), \\ & \text{Filme}(z_4, y, \text{“George Clooney”}), \\ & \text{Filme}(z_5, \text{“George Clooney”}, y) \end{aligned}$$

Unser Beweis von Lemma 3.14 übersetzt diese Anfrage in die Tableau-Anfrage $Q = (\mathbb{T}, u)$ mit Zusammenfassungstupel $u = (x, y)$ und folgendem Tableau \mathbb{T} :

<i>Programm</i>	Kino	Titel	Zeit
	z_1	x	z_2

<i>Filme</i>	Titel	Regie	Schauspieler
	x	y	z_3
	z_4	y	"George Clooney"
	z_5	"George Clooney"	y

Diese Tableau-Anfrage wiederum wird durch unseren Beweis übersetzt in die folgende Anfrage des konjunktiven Kalküls:

$$\left\{ \begin{array}{l} (x, y) : \exists z_1 \cdots \exists z_5 (\text{Programm}(z_1, x, z_2) \wedge \\ \text{Filme}(x, y, z_3) \wedge \\ \text{Filme}(z_4, y, \text{"George Clooney"}) \wedge \\ \text{Filme}(z_5, \text{"George Clooney"}, y)) \end{array} \right\}$$

Diese Anfrage ist in Normalform und wird durch unseren Beweis übersetzt in die regelbasierte Anfrage, mit der wir dieses Beispiel begonnen haben.

Folie 58

Die Kalkül-Anfrage

$$\left\{ \begin{array}{l} (x, y) : \exists z \exists z' \text{ Programm}(z, x, z') \wedge \\ \exists z \text{ Filme}(x, y, z) \wedge \\ \exists z \text{ Filme}(z, y, \text{"George Clooney"}) \wedge \\ \exists z \text{ Filme}(z_5, \text{"George Clooney"}, y) \end{array} \right\}$$

die nicht in Normalform ist, wird durch unseren Beweis zunächst in die Normalform-Anfrage

$$\left\{ \begin{array}{l} (x, y) : \exists z_1 \cdots \exists z_5 (\text{Programm}(z_1, x, z_2) \wedge \\ \text{Filme}(x, y, z_3) \wedge \\ \text{Filme}(z_4, y, \text{"George Clooney"}) \wedge \\ \text{Filme}(z_5, \text{"George Clooney"}, y)) \end{array} \right\}$$

übersetzt, die dann wiederum in die regelbasierte Anfrage übersetzt wird, mit der wir dieses Beispiel begonnen hatten.

Zur Info: Die im obigen Beispiel betrachtete Anfrage fragt nach Titel und Regisseur von zur Zeit laufenden Filmen, deren Regisseur sowohl als Schauspieler als auch als Regisseur schon mal mit George Clooney gearbeitet hat.

Folie 59

Einige Erweiterungen der Anfragesprachen

- (1) Test auf „Gleichheit“ von Variablen zulassen
- (2) Hintereinanderausführung (*Komposition*) mehrerer Anfrage zulassen

Zu (1):

Zum Beispiel lässt sich die Anfrage

- (6): Welche (je 2) Schauspieler haben schon in Filmen gespielt, in denen der jeweils andere Regie geführt hat?

ausdrücken durch

$$Ans(y_1, y_2) \leftarrow Filme(x_1, y_1, z_1), Filme(x_2, y_2, z_2), y_1=z_2, y_2=z_1$$

aber auch, äquivalent, durch

$$Ans(y_1, y_2) \leftarrow Filme(x_1, y_1, y_2), Filme(x_2, y_2, y_1)$$

Folie 60

Regelbasierte konjunktive Anfragen mit „=“

Prinzipiell:

Im Rumpf von Regeln auch Atome der Form „ $x=y$ “ und „ $x=c$ “ zulassen, für beliebige Variablen $x, y \in \mathbf{var}$ und Konstanten $c \in \mathbf{dom}$.

\leadsto regelbasierte konjunktive Anfragen mit „=“

\leadsto Konjunktiver Kalkül mit „=“

Aber Vorsicht: Die Regel Q der Form

$$Ans(x, y) \leftarrow R(x), y=z$$

ausgewertet in einer Datenbank \mathbf{I} mit $\mathbf{I}(R) = \{a\}$, liefert als Ergebnis

$$\llbracket Q \rrbracket(\mathbf{I}) = \{ (a, d) : d \in \mathbf{dom} \} = \{a\} \times \mathbf{dom}$$

Da \mathbf{dom} *unendlich viele Elemente* hat, ist $\llbracket Q \rrbracket(\mathbf{I})$ also eine „unendliche Relation“; per Definition, sind *als Relationen aber nur endliche Mengen erlaubt*.

Daher machen wir eine syntaktische Einschränkung auf *bereichsbeschränkte* Anfragen, um zu garantieren, dass das Ergebnis einer Anfrage stets eine endliche Relation ist.

Folie 61

Bereichsbeschränkte konjunktive Anfragen mit „=“

Präzise:

Jede Variable x , die im Rumpf der Regel vorkommt, muss auch in einem im Rumpf der Regel stehenden Atom der Form $R(u)$ oder $x=c$, für ein $c \in \mathbf{dom}$, vorkommen — oder es muss eine Kette von Gleichheits-Atomen der Form $x=y_1, y_1=y_2, \dots, y_j=z$ und ein Atom der Form $z=c$ für eine Konstante $c \in \mathbf{dom}$ geben oder ein Atom der Form $R(u)$ im Rumpf der Regel geben, so dass die Variable z im freien Tupel u vorkommt.

Regelbasierte konjunktive Anfragen mit „=“, die diese Bedingung erfüllen, nennen wir *bereichsbeschränkt*.

Analog wird die Klasse $\text{CQ}^=$ aller *bereichsbeschränkten Formeln des konjunktiven Kalküls mit „=“* definiert.

Beobachtung 3.15. *Jede $\text{CQ}^=$ -Formel ist entweder unerfüllbar oder äquivalent zu einer CQ-Formel. (Details siehe Übung.)*

Beispiel für eine $\text{CQ}^=$ -Anfrage, die nicht erfüllbar ist:

$$\{ (x) : (R(x) \wedge x=a \wedge x=b) \}$$

wobei a und b zwei verschiedene Elemente aus \mathbf{dom} sind.

Folie 62

Hintereinanderausführung mehrerer Anfragen

Im Folgenden widmen wir uns einer Erweiterung der konjunktiven Anfragen, die die Komposition mehrerer Anfragen ermöglicht, so dass eine Anfrage auf das Resultat einer (oder mehrerer) Anfragen angewendet werden kann.

Folie 63

Regelbasierte konjunktive Programme

Definition 3.16. Sei \mathbf{S} ein Datenbankschema.

Ein *regelbasiertes konjunktives Programm* über \mathbf{S} (mit oder ohne „=“) hat die Form

$$\begin{array}{lcl} S_1(u_1) & \leftarrow & \text{Rumpf}_1 \\ S_2(u_2) & \leftarrow & \text{Rumpf}_2 \\ \vdots & \vdots & \vdots \\ S_m(u_m) & \leftarrow & \text{Rumpf}_m \end{array}$$

wobei $m \geq 1$, S_1, \dots, S_m sind paarweise verschiedene Relationsnamen aus $\mathbf{rel} \setminus \mathbf{S}$ und für jedes $i \in \{1, \dots, m\}$ gilt:

$$Q_i := S_i(u_i) \leftarrow \text{Rumpf}_i$$

ist eine regelbasierte konjunktive Anfrage über $(\mathbf{S} \cup \{S_j : 1 \leq j < i\})$ (mit oder ohne „=“).

Die Relationsnamen aus \mathbf{S} , die im Programm vorkommen, heißen extensionale Prädikate (*edb-Prädikate*). Die Relationsnamen S_1, \dots, S_m heißen intensionale Prädikate (*idb-Prädikate*).

Folie 64

Semantik regelbasierter konjunktiver Programme

Ausgewertet über einer Datenbank $\mathbf{I} \in \text{inst}(\mathbf{S})$ beschreibt das obige Programm P die Relationen

$$\llbracket P(S_1) \rrbracket(\mathbf{I}), \dots, \llbracket P(S_m) \rrbracket(\mathbf{I}),$$

die induktiv für alle $i \in \{1, \dots, m\}$ wie folgt definiert sind:

$$\llbracket P(S_i) \rrbracket(\mathbf{I}) := \llbracket Q_i \rrbracket(\mathbf{J}_{i-1})$$

wobei \mathbf{J}_{i-1} die Erweiterung der Datenbank \mathbf{I} um die Relationen $\mathbf{J}(S_j) := \llbracket P(S_j) \rrbracket(\mathbf{I})$, für alle j mit $1 \leq j < i$ ist.

Folie 65

Äquivalenz von Regeln und Programmen

Beobachtung 3.17. Für jedes regelbasierte konjunktive Programm P über einem Relationenschema \mathbf{S} (mit oder ohne „=“) und jedes idb-Prädikat S von P gibt es eine regelbasierte konjunktive Anfrage Q (mit „=“) über \mathbf{S} , so dass

$$\llbracket Q \rrbracket(\mathbf{I}) = \llbracket P(S) \rrbracket(\mathbf{I})$$

für alle $\mathbf{I} \in \text{inst}(\mathbf{S})$.

Beispiel:

Sei $\mathbf{S} = \{Q, R\}$. Betrachte folgendes regelbasierte konjunktive Programm P :

$$\begin{aligned} S_1(x, z) &\leftarrow Q(x, y), R(y, z, w) \\ S_2(x, y, z) &\leftarrow S_1(x, w), R(w, y, v), S_1(v, z) \end{aligned}$$

Die von P durch S_2 definierte Anfrage ist äquivalent zu

$$S_2(x, y, z) \leftarrow \begin{aligned} &x=x', w=z', Q(x', y'), R(y', z', w'), \\ &R(w, y, v), \\ &v=x'', z=z'', Q(x'', y''), R(y'', z'', w'') \end{aligned}$$

3.2 Auswertungskomplexität

Folie 66

Auswertungskomplexität konjunktiver Anfragen

Auswertungsproblem für CQ (kombinierte Komplexität)
Eingabe: Anfrage Q des konjunktiven Kalküls,
 Datenbank \mathbf{I} (von einem zu Q passenden Schema \mathbf{S})
Aufgabe: Berechne $\llbracket Q \rrbracket(\mathbf{I})$

Schön wäre: Algorithmus, der zur Lösung dieses Problems mit Zeit polynomiell in „Größe der Eingabe + Größe der Ausgabe“ auskommt.

Frage: Gibt es einen solchen Algorithmus?

Größe der Eingabe: $k + n$, wobei

- $k := \|Q\|$ die Länge der Anfrage (betrachtet als Wort über dem Alphabet $\mathbf{dom} \cup \mathbf{var} \cup \mathbf{S} \cup \{\exists, \wedge, (,), \{, \}, :, , \}$)

Die Länge von konjunktiven regelbasierten Anfragen bzw. von Anfragen anderer Anfragesprachen wird analog definiert.

- $n := \|\mathbf{I}\|$ die Größe der Datenbank, also

$$n := \|\mathbf{I}\| := \sum_{R \in \mathbf{S}} \|\mathbf{I}(R)\| \quad \text{wobei} \quad \|\mathbf{I}(R)\| := \text{ar}(R) \cdot \underbrace{|\mathbf{I}(R)|}_{\text{Anzahl Tupel in } \mathbf{I}(R)}$$

Größe der Ausgabe:

$\|\llbracket Q \rrbracket(\mathbf{I})\| :=$ „Stelligkeit“ \cdot „Anzahl Tupel im Ergebnis“

Folie 67

Auswertung konjunktiver Anfragen

Proposition 3.18.

Das Auswertungsproblem für CQ lässt sich in Zeit $\mathcal{O}((k+n)^k)$ lösen.

Beweis: siehe unten.

Bemerkung: Das ist exponentiell in der Länge der Anfrage.

Frage: Geht das effizienter?

Beweis von Proposition 3.18:

Als Eingabe erhalten wir eine Datenbank \mathbf{I} und eine Anfrage Q der Form

$$\{ (e_1, \dots, e_r) : \varphi(x_1, \dots, x_m) \},$$

wobei (e_1, \dots, e_r) ein freies Tupel ist, dessen Variablen genau die Variablen aus $\{x_1, \dots, x_m\}$ sind, und φ eine CQ-Formel mit $\text{frei}(\varphi) = \{x_1, \dots, x_m\}$, die o.B.d.A. von der Form

$$\exists x_{m+1} \cdots \exists x_s \bigwedge_{i=1}^{\ell} R_i(u_i)$$

ist, wobei jedes u_i ein freies Tupel mit Variablen aus $\{x_1, \dots, x_s\}$ ist.

Unser Algorithmus geht bei Eingabe von Q und \mathbf{I} wie folgt vor:

- (1) Bestimme $\text{adom}(Q, \mathbf{I})$.
- (2) Für jedes $\bar{b} = (b_1, \dots, b_s) \in (\text{adom}(Q, \mathbf{I}))^s$ tue Folgendes:
 - (2.1) Sei $\beta : \{x_1, \dots, x_s\} \rightarrow \mathbf{dom}$ die Belegung mit $\beta(x_i) = b_i$ für jedes $i \in \{1, \dots, s\}$.
 - (2.2) **test** := **true**
 - (2.3) Für $i = 1, \dots, \ell$ tue Folgendes:
 - (2.3.1) Teste, ob $\beta(u_i) \in \mathbf{I}(R_i)$ gilt.
 - (2.3.2) Falls nicht, so setze **test** := **false**.
 - (2.4) Falls **test** = **true**, so gib das Tupel $(\beta(e_1), \dots, \beta(e_r))$ aus.

Unter Verwendung von Proposition 3.10 („ $\text{adom}(\llbracket Q \rrbracket(\mathbf{I})) \subseteq \text{adom}(Q, \mathbf{I})$ “) sieht man leicht, dass dieser Algorithmus genau die Tupel aus $\llbracket Q \rrbracket(\mathbf{I})$ ausgibt.

Eine Laufzeitanalyse für $k := \|Q\|$, $n := \|I\|$ und $N := k + n$ ergibt, dass

- für Zeile (1) nur $\mathcal{O}(N)$ Schritte gemacht werden (wir verzichten hier darauf, Duplikate zu eliminieren und erstellen einfach eine Liste aller beim Lesen von Q und \mathbf{I} gefundenen Konstanten),
- die Schleife in Zeile (2) höchstens N^s mal durchlaufen wird und
- bei jedem Schleifendurchlauf höchstens $s + 1 + \ell \cdot (n + 1) + 1 + r$ Schritte gemacht werden.

Eine nähere Betrachtung von Q zeigt, dass

$$k = \|Q\| \geq 2 + 2r + 1 + 2(s-m) + \ell \cdot 4$$

ist. Außerdem ist $r \geq m$, und daher ist $r + (s-m) \geq s$, und somit ist

$$k \geq 3 + r + (s-m) + s + 4\ell \geq s + 1 + \ell + 1 + r.$$

Die Gesamtlaufzeit unseres Algorithmus ist also

$$\begin{aligned} \mathcal{O}(N) + N^s \cdot (s + 1 + \ell + 1 + r + \ell \cdot n) &\leq \mathcal{O}(N) + N^s \cdot (k + \ell \cdot n) \\ &\leq \mathcal{O}(N) + N^s \cdot N^2 \\ &\leq \mathcal{O}(N^{s+2}) \\ &\leq \mathcal{O}(N^k) \\ &\leq \mathcal{O}((k+n)^k). \end{aligned}$$

Dies beendet den Beweis von Proposition 3.18. □

Folie 68

Boolesche Anfragen

- *Zur Erinnerung:*
Boolesche Anfragen sind „ja / nein“-Anfragen, d.h. Anfragen, deren Ergebnis die Stelligkeit 0 hat.
- *Klar:*
Falls wir zeigen können, dass das Auswertungsproblem für Boolesche Anfragen des konjunktiven Kalküls schwierig ist, so ist es auch für allgemeine Anfragen des konjunktiven Kalküls schwierig.
- *Wir werden sehen, dass umgekehrt aber auch gilt:*
Falls wir einen Algorithmus haben, der das Auswertungsproblem für *Boolesche* Anfragen des konjunktiven Kalküls löst, so können wir diesen Algorithmus verwenden, um das Auswertungsproblem für *beliebige* Anfragen des konjunktiven Kalküls zu lösen.

Folie 69

Algorithmen mit Verzögerung $f(k, n)$

Definition 3.19. Sei $f : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}$, sei \mathcal{A} eine Datenbankanfragesprache und sei \mathbb{A} irgendein Algorithmus, der bei allen Eingaben terminiert.

Wir sagen:

das Auswertungsproblem für \mathcal{A} kann unter Rückgriff auf \mathbb{A} mit Verzögerung $f(k, n)$ gelöst werden,

falls es einen Algorithmus \mathbb{B} gibt, der bei Eingabe einer Anfrage Q aus \mathcal{A} und einer Datenbank \mathbf{I} nach und nach genau die Tupel aus $\llbracket Q \rrbracket(\mathbf{I})$ ausgibt und

- vor der Ausgabe des ersten Tupels,
- zwischen der Ausgabe von je zwei aufeinanderfolgenden Tupeln,
- nach der Ausgabe des letzten Tupels

je höchstens $f(\|Q\|, \|\mathbf{I}\|)$ viele Elementarschritte oder Schritte, in denen der Algorithmus \mathbb{A} aufgerufen wird, macht.

Folie 70

Boolesche Anfragen \rightsquigarrow beliebige Anfragen

Theorem 3.20. Sei \mathbb{A} ein Algorithmus, der das Auswertungsproblem für Boolesche Anfragen des konjunktiven Kalküls löst.

Dann gibt es einen Algorithmus \mathbb{B} , der das Auswertungsproblem für (beliebige) Anfragen des konjunktiven Kalküls unter Rückgriff auf \mathbb{A} mit Verzögerung $\mathcal{O}(k^3 \cdot n)$ löst.

Beweis: siehe unten.

Folgerung: Falls wir das Auswertungsproblem für Boolesche Anfragen des konjunktiven Kalküls effizient lösen können, dann können wir es auch für beliebige Anfragen des konjunktiven Kalküls effizient lösen.

Beweis von Theorem 3.20:

Wir betrachten o.B.d.A. nur Anfragen der Form

$$\{ (x_1, \dots, x_r, c_1, \dots, c_m) : \varphi(x_1, \dots, x_r) \}$$

wobei für $r, m, s, \ell \geq 0$ gilt:

- x_1, \dots, x_r sind paarweise verschiedene Variablen.
Die Zahl r wird im Folgenden *Stelligkeit der Anfrage* genannt.
- c_1, \dots, c_m sind (nicht notwendigerweise unterschiedliche) Konstanten.
- $\text{frei}(\varphi) = \{x_1, \dots, x_r\}$.
- φ ist von der Form $\exists x_{r+1} \cdots \exists x_s \bigwedge_{i=1}^{\ell} R_i(u_i)$, wobei jedes u_i ein freies Tupel mit Variablen aus $\{x_1, \dots, x_s\}$ ist.

Bei Eingabe einer Datenbank \mathbf{I} und einer Anfrage Q dieser Form geht unser Algorithmus \mathbb{B} in 2 Phasen vor.

Phase 1: Ermittle die Stelligkeit r der Anfrage und erstelle eine Liste a_1, \dots, a_N aller Elemente aus $\text{adom}(Q, \mathbf{I})$.

Für $k := \|Q\|$ und $n := \|\mathbf{I}\|$ gilt $N \leq k + n$. Und Phase 1 benötigt nur Zeit $\mathcal{O}(k + n)$ (wir verzichten hier darauf, Duplikate zu eliminieren und erstellen einfach eine Liste aller beim Lesen von Q und \mathbf{I} gefundenen Konstanten).

Phase 2: Berechne $\llbracket Q \rrbracket(\mathbf{I})$.

Unter Verwendung von Proposition 3.10 erhalten wir, dass Folgendes gilt:

$$\llbracket Q \rrbracket(\mathbf{I}) \subseteq \{a_1, \dots, a_N\}^r \times \{c_1\} \times \cdots \times \{c_m\}.$$

Eine erste und noch nicht ganz so gute Idee, $\llbracket Q \rrbracket(\mathbf{I})$ zu berechnen, ist wie folgt vorzugehen.

Für alle $b_r \in \{a_1, \dots, a_N\}$ tue Folgendes:

 Für alle $b_{r-1} \in \{a_1, \dots, a_N\}$ tue Folgendes:

\dots

 Für alle $b_1 \in \{a_1, \dots, a_N\}$ tue Folgendes:

 (1) Starte Algorithmus \mathbb{A} mit DB \mathbf{I} und Anfrage

$$Q_{(b_1, \dots, b_r)} := \{ () : \varphi(x_1/b_1, \dots, x_r/b_r) \}$$

 (2) Falls \mathbb{A} „ja“ ausgibt, so gib das Tupel

$$(b_1, \dots, b_r, c_1, \dots, c_m) \text{ aus.}$$

Man kann sich leicht davon überzeugen, dass dieser Algorithmus genau die Tupel aus $\llbracket Q \rrbracket(\mathbf{I})$ ausgibt. Das Problem ist allerdings, dass vor der Ausgabe des ersten Tupels (und zwischen der Ausgabe zweier Tupel und nach der Ausgabe des letzten Tupels) evtl. zu viel Zeit vergeht. Wenn beispielsweise $(a_N, \dots, a_N, c_1, \dots, c_m)$ das *einzig*e Tupel in $\llbracket Q \rrbracket(\mathbf{I})$ ist, dann gibt der obige Algorithmus dieses erst im allerletzten Schleifendurchlauf aus, also nach

etwa N^r Schritten. Das „ r “, das hier im Exponenten steht, hängt von der Anfrage Q ab. Die hier erzielte Verzögerung von N^r ist u.U. also viel größer als die im Theorem behauptete Verzögerung von maximal $k^3 \cdot n$.

Lösungsansatz: Die Abarbeitung der ineinandergeschachtelten Schleifen des obigen Algorithmus können wir uns als einen Tiefensuche-Durchlauf durch den folgenden *Suchbaum* vorstellen: Der Suchbaum ist ein vollständiger Baum der Tiefe r , bei dem jeder Knoten, der sich auf einer Tiefe $< r$ befindet, genau N Kinder hat (und diese sind mit a_1, \dots, a_N beschriftet). Die N Kinder v_1, \dots, v_N der Wurzel stellen alle möglichen Belegungen für b_r dar; jedes v_i hat wieder N Kinder, die alle möglichen Belegungen für b_{r-1} darstellen, usw.

Unser neuer Lösungsansatz soll nun sicherstellen, dass nur solche Kanten des Suchbaums durchlaufen werden, die tatsächlich zu einem Tupel aus $\llbracket Q \rrbracket(\mathbf{I})$ führen. Dies kann man dadurch bewerkstelligen, dass man vor dem Durchlaufen einer Kante eine Boolesche Anfrage stellt, die nachfragt, ob es im Teilbaum unterhalb dieser Kante ein Tupel gibt, das zu $\llbracket Q \rrbracket(\mathbf{I})$ führt. Dies wird erreicht, indem wir die folgende Prozedur \mathbb{C} mit Eingabe r, Q starten.

Prozedur $\mathbb{C}(r, Q)$:

Falls $r = 0$:

- (1) Starte Algorithmus \mathbb{A} mit DB \mathbf{I} und Boolescher Anfrage $Q^{\text{Bool}} := \{ () : \varphi \}$
- (2) Falls \mathbb{A} „ja“ ausgibt, so gib das Tupel (c_1, \dots, c_m) aus und halte an.

Falls $r > 0$:

Für jedes $b_r \in \{a_1, \dots, a_N\}$ tue Folgendes:

- (3) Starte Algorithmus \mathbb{A} mit DB \mathbf{I} und Boolescher Anfrage $Q^{\text{ex.Lsg}} := \{ () : \exists x_1 \dots \exists x_{r-1} \varphi(x_1, \dots, x_{r-1}, x_r/b_r) \}$
- (4) Falls \mathbb{A} „ja“ ausgibt, so starte einen rekursiven Aufruf der Prozedur \mathbb{C} mit Eingabe r' und Q' , für
 - $r' := r - 1$
 - $Q' := Q_{b_r}^{\text{berechne Lsgen}} := \{ (x_1, \dots, x_{r-1}, b_r, c_1, \dots, c_m) : \varphi(x_1, \dots, x_{r-1}, x_r/b_r) \}$

Per Induktion nach r zeigen wir, dass Folgendes gilt:

- (1)_r: $\mathbb{C}(r, Q)$ gibt genau die Tupel aus $\llbracket Q \rrbracket(\mathbf{I})$ aus.
- (2)_r: $\mathbb{C}(r, Q)$ hat Verzögerung höchstens $(r+1) \cdot (2k+1) \cdot (N+1)$, d.h. vor der Ausgabe des ersten Tupels, zwischen der Ausgabe zweier Tupel und nach der Ausgabe des letzten Tupels führt die Prozedur jeweils höchstens $(r+1) \cdot (2k+1) \cdot (N+1)$ Elementarschritte oder Aufrufe von \mathbb{A} durch.

Induktionsanfang für $r = 0$:

Aussage (1)₀ gilt offensichtlich.

Außerdem macht die Prozedur maximal k Elementarschritte, um die Anfrage Q^{Bool} zu konstruieren, startet danach einen Aufruf von \mathbb{A} und macht ggf. nochmal m Elementarschritte, um das Tupel (c_1, \dots, c_m) auszugeben. Die Gesamtzahl der Schritte ist also höchstens $k + 1 + m \leq 2k + 1 \leq (r+1)(2k+1)(N+1)$. Somit ist Aussage (2)₀ erfüllt.

Induktionsschritt für $r > 0$:

Die Gültigkeit der Aussage (1)_r folgt direkt aus der Konstruktion der Prozedur \mathbb{C} , der Induktionsannahme und der Tatsache, dass $\text{adom}(\llbracket Q \rrbracket(\mathbf{I})) \subseteq \text{adom}(Q, \mathbf{I})$ ist (Details: Übung).

Zum Nachweis der Gültigkeit der Aussage (2)_r beachte Folgendes:

Vor der Ausgabe des ersten Tupels, zwischen der Ausgabe von zwei Tupeln und nach der Ausgabe des letzten Tupels macht die Prozedur je höchstens

- (A) N Schritte, in denen Zeile (3) aufgerufen wird und die Ausgabe „nein“ liefert.

Bei jedem solchen Schritt fallen maximal k Elementarschritte zur Konstruktion der Anfrage $Q^{\text{ex.Lsg}}$ an, und 1 Schritt zum Aufruf von \mathbb{A} .

- (B) 1 Schritt, in dem Zeile (3) aufgerufen wird und die Ausgabe „ja“ liefert, so dass danach Zeile (4) aufgerufen wird (und wir wissen, dass während der Abarbeitung dieses Aufrufs mindestens ein Tupel ausgegeben wird).

Dabei werden höchstens folgende Schritte gemacht:

- k Elementarschritte zur Konstruktion von $Q_{b_r}^{\text{ex.Lsg}}$
- 1 Schritt zum Aufruf von \mathbb{A}
- k Elementarschritte zur Konstruktion von $Q_{b_r}^{\text{berechne Lsgen}}$
- nach dem Start von $\mathbb{C}(r-1, Q_{b_r}^{\text{berechne Lsgen}})$ gemäß Induktionsannahme maximal $r \cdot (2k+1) \cdot (N+1)$ Schritte bis zur Ausgabe des ersten Tupels, zwischen der Ausgabe von 2 Tupeln und nach der Ausgabe des letzten Tupels.

Bei (A) werden also insgesamt höchstens $(k+1) \cdot N$ Schritte gemacht, und bei (B) beträgt die „Verzögerung“ höchstens $2k + 1 + r(2k+1)(N+1)$. Insgesamt haben wir also eine „Verzögerung“ von höchstens

$$\begin{aligned} & (k+1)N + 2k + 1 + r \cdot (2k+1) \cdot (N+1) \\ \leq & (2k+1) \cdot (N+1) + r \cdot (2k+1) \cdot (N+1) \\ \leq & (r+1) \cdot (2k+1) \cdot (N+1) \end{aligned}$$

Schritten. Somit ist also Aussage $(2)_r$ bewiesen.

Insgesamt erhalten wir, dass unser Algorithmus \mathbb{B} das Auswertungsproblem unter Rückgriff auf \mathbb{A} mit Verzögerung

$$\mathcal{O}(\underbrace{k+n}_{\text{Phase 1}} + \underbrace{r \cdot k \cdot N}_{\text{Phase 2}}) \leq \mathcal{O}(k^2 \cdot (k+n)) \leq \mathcal{O}(k^3 \cdot n)$$

löst. Dies beendet den Beweis von Theorem 3.20. □

Folie 71

Die Komplexitätsklasse NP

Zur Erinnerung:

- Komplexitätsklassen bestehen aus *Entscheidungsproblemen*, d.h. Problemen mit „ja/nein“-Ausgabe
 - \rightsquigarrow das *Auswertungsproblem für Boolesche Anfragen* passt gut zum Konzept der klassischen Komplexitätstheorie;
 - \rightsquigarrow das Auswertungsproblem für beliebige Anfragen nicht
- Ein Entscheidungsproblem B gehört zur Klasse NP, falls es einen nichtdeterministischen Algorithmus \mathbb{B} gibt, der eine *durch ein Polynom beschränkte worst-case Laufzeit* besitzt und der bei Eingabe jeder für B zulässigen Eingabe x Folgendes leistet:
 - Falls x eine „ja“-Instanz von B ist, so besitzt \mathbb{B} bei Eingabe x mindestens einen Berechnungspfad, der mit der Ausgabe „ja“ endet.
 - Falls x eine „nein“-Instanz von B ist, so endet *jeder* Berechnungspfad von \mathbb{B} mit der Ausgabe „nein“.

„Nichtdeterministische Algorithmen“ werden in der Komplexitätstheorie i.d.R. durch nichtdeterministische Turingmaschinen modelliert; ein „Berechnungspfad“ entspricht dann einem *Lauf* der Turingmaschine.

Folie 72

NP-Vollständigkeit

Zur Erinnerung:

Ein Entscheidungsproblem B heißt *NP-vollständig* (bzgl. Polynomialzeit-Reduktionen), falls gilt:

- (1) $B \in \text{NP}$ und
- (2) B ist *NP-hart*, d.h. für jedes Problem $A \in \text{NP}$ gibt es eine *Polynomialzeit-Reduktion* f von A auf B (kurz: $f : A \leq_p B$)

Folie 73

Exkurs: Reduktionen

- Wir wollen einen Algorithmus \mathbb{A} finden, der ein Entscheidungsproblem A löst.

Beispiel: A ist das Problem, bei Eingabe eines Graphen G und einer Zahl k zu entscheiden, ob G eine Clique der Größe k besitzt.¹

Mit M_A bezeichnen wir im Folgenden die Menge aller für das Problem A zulässigen Eingaben.

- Nehmen wir mal an, wir hätten bereits einen Algorithmus \mathbb{B} konstruiert, der ein Entscheidungsproblem B löst.

Beispiel: B ist das Problem, bei Eingabe eines Graphen H und einer Zahl ℓ zu entscheiden, ob H eine unabhängige Menge der Größe ℓ besitzt.²

¹Eine Clique der Größe k in $G = (V, E)$ ist eine Menge $C \subseteq V$ mit $|C| = k$, so dass für alle $u, v \in C$ mit $u \neq v$ gilt: u und v sind in G durch eine Kante verbunden.

²Eine unabhängige Menge der Größe ℓ in $H = (V, E)$ ist eine Menge $U \subseteq V$ mit $|U| = \ell$, so dass für alle $u, v \in U$ mit $u \neq v$ gilt: u und v sind in H *nicht* durch eine Kante verbunden.

Mit M_B bezeichnen wir die Menge aller für das Problem B zulässigen Eingaben.

Folie 74

- **Idee zur Konstruktion von \mathbb{A} :** Nutze den Algorithmus \mathbb{B} .

Wir bekommen eine für A zulässige Eingabe x und wollen entscheiden, ob x eine „ja“-Instanz für A ist.

Finde eine **effizient berechenbare** Funktion $f : M_A \rightarrow M_B$, so dass für alle Eingaben x für A gilt:

$$x \text{ ist eine „ja“-Instanz für das Problem } A \iff f(x) \text{ ist eine „ja“-Instanz für das Problem } B$$

Dann kann Algorithmus \mathbb{A} bei Eingabe x wie folgt vorgehen:

1. Berechne $y := f(x)$,
2. starte Algorithmus \mathbb{B} mit Eingabe y ,
3. gib das aus, was \mathbb{B} ausgibt.

Beispiel:

$f(G, k) = (\overline{G}, k)$, wobei \overline{G} das **Komplement** des Graphen G ist.

- f heißt **Reduktion** von A auf B . Kurz: $f : A \leq B$.

Folie 75

- Eine Reduktion f von A auf B heißt *Polynomialzeit-Reduktion* (kurz: $f : A \leq_p B$), falls es einen (deterministischen) Polynomialzeit-Algorithmus gibt, der bei Eingabe von $x \in M_A$ den Wert $f(x)$ berechnet.
- Wenn es eine Polynomialzeit-Reduktion f von A auf B gibt, wissen wir, dass Folgendes gilt:
 - Ein Algorithmus, der B löst, kann zum Lösen von A verwendet werden. Das Problem A ist also „höchstens so schwer“ wie das Problem B .
 - Das heißt umgekehrt aber auch, dass das Problem B „mindestens so schwer“ wie das Problem A ist.

Folie 76

NP-Vollständigkeit

Zur Erinnerung:

- Ein Entscheidungsproblem B heißt *NP-vollständig* (bzgl. Polynomialzeit-Reduktionen), falls gilt:
 - (1) $B \in \text{NP}$ und
 - (2) B ist *NP-hart*, d.h. für jedes Problem $A \in \text{NP}$ gibt es eine *Polynomialzeit-Reduktion* f von A auf B (kurz: $f : A \leq_p B$)
- Eine *Polynomialzeit-Reduktion* f von A auf B ist eine (deterministisch) in polynomieller Zeit berechenbare Funktion, die jede zum Problem A passende Eingabe x auf eine zum Problem B passende Eingaben $f(x)$ abbildet, so dass gilt

x ist eine „ja“-Instanz für $A \iff f(x)$ ist eine „ja“-Instanz für B .

Anschaulich bedeutet $A \leq_p B$, dass A „höchstens so schwer“ wie B ist. NP-vollständige Probleme sind also „die schwierigsten Probleme“ in NP.
- Der Begriff *Polynomialzeit-Reduktion* ist so definiert, dass Folgendes gilt, wobei P die Klasse aller Entscheidungsprobleme bezeichnet, die deterministisch in Polynomialzeit lösbar sind:
 - $A \leq_p B$ und $B \in P \implies A \in P$
 - $A \notin P$ und $A \leq_p B \implies B \notin P$
 - A NP-hart und $A \leq_p B \implies B$ NP-hart.

Folie 77

Auswertungskomplexität konjunktiver Anfragen

Theorem 3.21 (Chandra, Merlin, 1977). *Das Auswertungsproblem (kombinierte Komplexität) für Boolesche regelbasierte konjunktive Anfragen ist NP-vollständig.*

Beweis:

Zugehörigkeit zu NP: Bei Eingabe einer DB I und einer Booleschen Anfrage Q der Form

$$\text{Ans}() \leftarrow R_1(u_1), \dots, R_\ell(u_\ell)$$

mit $\text{Var}(Q) = \{x_1, \dots, x_s\}$ kann ein nichtdeterministischer Algorithmus wie folgt vorgehen:

- (1) Bestimme $\text{adom}(Q, \mathbf{I})$
- (2) „Rate“ (nichtdeterministisch) Elemente b_1, \dots, b_s aus $\text{adom}(Q, \mathbf{I})$.
Sei $\beta : \text{Var}(Q) \rightarrow \mathbf{dom}$ die Belegung mit $\beta(x_j) := b_j$ f.a. $j \in \{1, \dots, s\}$.
- (3) Teste, ob für alle $i \in \{1, \dots, \ell\}$ gilt: $\beta(u_i) \in \mathbf{I}(R_i)$.
- (4) Falls „ja“, so STOPP mit Ausgabe „ja“;
ansonsten STOPP mit Ausgabe „nein“.

Man kann leicht nachprüfen, dass dies ein nichtdeterministischer Algorithmus ist, der das Auswertungsproblem (kombinierte Komplexität) für Boolesche regelbasierte konjunktive Anfragen in Zeit $\mathcal{O}(k \cdot n)$ löst. Insbesondere liegt das Problem also in NP.

NP-Härte: Zum Nachweis der NP-Härte benutzen wir das folgende Resultat, das Sie bereits aus der Veranstaltung *Einführung in die Theoretische Informatik* kennen:

Theorem: CLIQUE ist NP-vollständig.

Das Problem CLIQUE ist dabei folgendermaßen definiert:

CLIQUE

Eingabe: Ein endlicher ungerichteter Graph $G = (V, E)$ und eine Zahl $k \in \mathbb{N}$

Frage: Enthält G eine k -Clique?

(D.h. gibt es k verschiedene Knoten in G , von denen jeder mit jedem anderen durch eine Kante verbunden ist?)

Bei einem *endlichen ungerichteten Graphen* $G = (V, E)$ ist jede Kante in E ist eine 2-elementige Teilmenge von V .

Wir beweisen die NP-Härte des Auswertungsproblems (kombinierte Komplexität) für regelbasierte konjunktive Anfragen, indem wir eine Polynomialzeit-Reduktion f vom CLIQUE-Problem auf dieses Problem angeben.

Sei (G, k) eine Eingabe für's CLIQUE-Problem, und sei $G = (V, E)$. Unsere Reduktion f bildet die CLIQUE-Instanz (G, k) auf die wie folgt definierte Instanz $f(G, k) := (Q_{(G,k)}, \mathbf{I}_G)$ des Auswertungsproblems ab:

- Sei $\mathbf{S} := \{R\}$ das Datenbankschema, das aus einem Relationsnamen R der Stelligkeit 2 besteht.

- Sei \mathbf{I}_G die Datenbank über dem Schema \mathbf{S} mit

$$\mathbf{I}_G(R) := \{ (u, v) : \{u, v\} \in E \}$$

- Sei $Q_{(G,k)}$ die regelbasierte konjunktive Anfrage

$$\text{Ans}() \leftarrow (R(x_i, x_j),)_{1 \leq i < j \leq k'}$$

wobei $k' := \min\{k, |V|+1\}$. Mit $(R(x_i, x_j),)_{1 \leq i < j \leq k'}$ ist hier gemeint, dass der Rumpf der Anfrage aus einer durch Kommas getrennten Liste aller Atome $R(x_i, x_j)$, für alle i und j mit $1 \leq i < j \leq k'$, bestehen soll.

Beispiel:

Ist $G = (V, E)$ der Graph mit Knotenmenge $V = \{a, b, c, d\}$, dessen Kantenmenge aus den Kanten $\{a, b\}$ und $\{a, c\}$ besteht, dann ist \mathbf{I}_G die Datenbank, deren Relation $\mathbf{I}_G(R)$ aus den Tupeln (a, b) , (b, a) , (a, c) und (c, a) besteht. Ist $k = 3$, so ist außerdem $k' = 3$, und $Q_{(G,k)}$ ist die Anfrage

$$\text{Ans}() \leftarrow R(x_1, x_2), R(x_1, x_3), R(x_2, x_3).$$

Man sieht leicht, dass bei Eingabe von (G, k) die Datenbank \mathbf{I}_G und die Anfrage $Q_{(G,k)}$ in Zeit polynomiell in der Größe von (G, k) erzeugt werden können. Dabei ist zu beachten, dass die Größe des Graphen $G = (V, E)$ definiert ist als $|V| + |E|$, während die „Größe der Zahl k “ definiert ist als die Anzahl der Bits, die nötig sind, um k zu repräsentieren — und das sind $\mathcal{O}(\log k)$ viele Bits. Den Wert k' haben wir so gewählt, dass die Anfrage $Q_{(G,k)}$ in Zeit polynomiell in $|V| + |E| + \log k$ erzeugt werden kann.

Die Funktion f kann also in Polynomialzeit berechnet werden. Um den Beweis abzuschließen, müssen wir nur noch beweisen, dass Folgendes gilt:

$$G \text{ besitzt eine } k\text{-Clique} \iff \llbracket Q_{(G,k)} \rrbracket(\mathbf{I}_G) = \{ () \}.$$

Für die Richtung „ \implies “ sei v_1, \dots, v_k eine Liste von k verschiedenen Knoten von G , die eine k -Clique bilden. Insbesondere gilt dann:

- $|V| \geq k$, und daher ist $k' = k$; und
- für alle $i, j \in \{1, \dots, k\}$ mit $i < j$ gibt es in E die Kante $\{v_i, v_j\}$.

Sei β die Belegung für $Q_{(G,k)}$ mit $\beta(x_i) := v_i$ für jedes $i \in \{1, \dots, k\}$. Gemäß unserer Wahl der Datenbank \mathbf{I}_G gilt dann für alle $i, j \in [k]$ mit $i < j$, dass $(\beta(x_i), \beta(x_j)) \in \mathbf{I}_G(R)$ ist. Somit bezeugt β , dass $\llbracket Q_{(G,k)} \rrbracket(\mathbf{I}_G) = \{ () \}$ ist.

Für die Richtung „ \Leftarrow “ sei $\beta : \{x_1, \dots, x_{k'}\} \rightarrow \mathbf{dom}$ eine Belegung für $Q_{(G,k)}$, die bezeugt, dass $\llbracket Q_{(G,k)} \rrbracket(\mathbf{I}_G) = \{ () \}$ ist. Also gilt für alle $i, j \in [k']$ mit $i < j$, dass $(\beta(x_i), \beta(x_j)) \in \mathbf{I}_G(R)$.

Für alle $i \in \{1, \dots, k'\}$ setze $v_i := \beta(x_i)$. Gemäß unserer Konstruktion der Datenbank \mathbf{I}_G folgt aus $(\beta(x_i), \beta(x_j)) \in \mathbf{I}_G(R)$, dass $\{v_i, v_j\} \in E$ ist — und da G ungerichtet ist, gilt insbesondere auch $v_i \neq v_j$.

Somit ist $v_1, \dots, v_{k'}$ eine Liste von k' verschiedenen Knoten, die in G eine k' -Clique bilden. Insbesondere ist $k' \leq |V|$, und daher ist

$k' \stackrel{\text{Def}}{=} \min\{k, |V|+1\} = k$. Der Graph G besitzt also eine k -Clique.

Dies beendet den Beweis von Theorem 3.21. □

Folie 78

Folgerung aus der NP-Vollständigkeit

Folgerung:

Falls $P \neq NP$, so gibt es keinen Algorithmus, der das Auswertungsproblem für Boolesche Anfragen des konjunktiven Kalküls deterministisch in Zeit $(k+n)^{\mathcal{O}(1)}$ löst.

Bemerkung 3.22 (hier ohne Beweis). Unter Verwendung einer stärkeren Annahme aus der *Parametrischen Komplexitätstheorie* lässt sich Folgendes zeigen:

Theorem (Papadimitriou, Yannakakis, 1997)

Falls $FPT \neq W[1]$, so gibt es keinen Algorithmus, der das Auswertungsproblem für Boolesche Anfragen des konjunktiven Kalküls deterministisch in Zeit $f(k) \cdot n^c$ löst (wobei f irgendeine berechenbare Funktion und c irgendeine Konstante ist).

FPT und $W[1]$ sind Komplexitätsklassen, die in der parametrischen Komplexitätstheorie Rollen spielen, die in etwa mit den Rollen von P und NP in der klassischen Komplexitätstheorie vergleichbar sind.

3.3 Algebraischer Ansatz: SPC-Algebra und SPJR-Algebra

Folie 79

Algebraische Anfragen — Informell

Beispiel-Anfrage:

(4) Welche Kinos (Name & Adresse) spielen einen Film mit “Matt Damon”?

Als Anfrage in der SPJR-Algebra:

$$\pi_{Kino, Adresse} \left(\sigma_{\substack{\text{Schauspieler} = \\ \text{“Matt Damon”}}} (Filme \bowtie Programm \bowtie \delta_{\text{Name} \mapsto \text{Kino}} Kinos) \right)$$

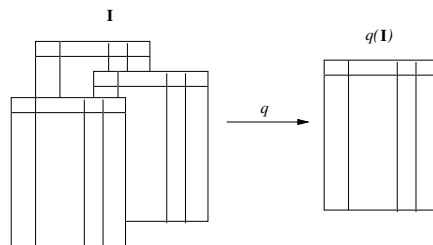
Als Anfrage in der SPC-Algebra:

$$\pi_{7,8} \left(\sigma_{4=7} \left(\sigma_{1=5} \left(\sigma_{3 = \text{“Matt Damon”}} (Filme \times Programm \times Kinos) \right) \right) \right)$$

Folie 80

SPC-Algebra bzw. SPJR-Algebra

Zur Erinnerung:



- Mathematik:
 - Algebraische Struktur $\hat{=}$ Grundmenge + Operationen
- Hier:
 - Operationen auf (endlichen) Relationen
 - Speziell: Projektion, Selektion, Kartesisches Produkt bzw. Join und Umbenennung

Folie 81

Unbenannte Perspektive: Die SPC-Algebra

Selektion: Zwei Varianten:

- *Operator* $\sigma_{j=a}$, für eine Konstante $a \in \mathbf{dom}$ und eine natürliche Zahl $j \geq 1$.
Dieser Operator kann angewendet werden auf Relationen I der Stelligkeit $\geq j$ und liefert als Ausgabe die folgende Teilmenge der Relation I :

$$\sigma_{j=a}(I) := \{t \in I : \text{in der } j\text{-ten Komponente von } t \text{ steht ein } a\}$$

- *Operator* $\sigma_{j=k}$, für zwei natürliche Zahlen $j, k \geq 1$.
Dieser Operator kann angewendet werden auf Relationen I der Stelligkeit $\geq \max(j, k)$ und liefert als Ausgabe die folgende Teilmenge der Relation I :

$$\sigma_{j=k}(I) := \left\{ t \in I : \begin{array}{l} \text{in der } j\text{-ten Komponente von } t \text{ steht derselbe} \\ \text{Eintrag wie in der } k\text{-ten Komponente von } t \end{array} \right\}$$

Selektion ist eine „horizontale“ Operation: sie wählt einzelne Tabellen-Zeilen aus.

„ $j=a$ “ und „ $j=k$ “ werden *Selektionsbedingungen* genannt.

Folie 82

Projektion:

Operator π_{j_1, \dots, j_k} , für (nicht notwendigerweise paarweise verschiedene) natürliche Zahlen $j_1, \dots, j_k \geq 1$ (und $k \geq 0$).

Dieser Operator kann angewendet werden auf Relationen I der Stelligkeit $\geq \max\{j_1, \dots, j_k\}$ und liefert als Ausgabe die folgende Relation der Stelligkeit k :

$$\pi_{j_1, \dots, j_k}(I) := \{ (t(j_1), \dots, t(j_k)) : t \in I \}$$

Projektion ist eine „vertikale“ Operation: sie wählt einzelne Tabellen-Spalten aus und arrangiert sie in möglicherweise anderer Reihenfolge

Folie 83

Kartesisches Produkt:

Operator \times

Dieser Operator kann angewendet werden auf Relationen I und J beliebiger Stelligkeiten m und n und liefert als Ausgabe die folgende Relation der Stelligkeit $m+n$:

$$I \times J := \left\{ \left(t(1), \dots, t(m), s(1), \dots, s(n) \right) : t \in I \text{ und } s \in J \right\}$$

Wir benutzen den Operator manchmal auch für einzelne Tupel: Sind t und s Tupel der Stelligkeiten m und n , so schreiben wir $t \times s$, um das Tupel $(t(1), \dots, t(m), s(1), \dots, s(n))$ zu bezeichnen.

Folie 84

Definition 3.23.

Sei \mathbf{S} ein Datenbankschema. Die Klasse der Anfragen der *SPC-Algebra über \mathbf{S}* (kurz: $\text{SPC}[\mathbf{S}]$) ist induktiv wie folgt definiert:

- Für alle Relationsnamen $R \in \mathbf{S}$ ist R eine $\text{SPC}[\mathbf{S}]$ -Anfrage der Stelligkeit $\text{ar}(R)$.
- Für alle Konstanten $c \in \mathbf{dom}$ ist $\{(c)\}$ eine $\text{SPC}[\mathbf{S}]$ -Anfrage der Stelligkeit 1.
- Ist Q eine $\text{SPC}[\mathbf{S}]$ -Anfrage der Stelligkeit m , sind j, k natürliche Zahlen aus $\{1, \dots, m\}$ und ist $a \in \mathbf{dom}$ eine Konstante, so sind auch $\sigma_{j=a}(Q)$ und $\sigma_{j=k}(Q)$ $\text{SPC}[\mathbf{S}]$ -Anfragen der Stelligkeit m .
- Ist Q eine $\text{SPC}[\mathbf{S}]$ -Anfrage der Stelligkeit m , ist $k \geq 0$ und sind j_1, \dots, j_k natürliche Zahlen aus $\{1, \dots, m\}$, so ist $\pi_{j_1, \dots, j_k}(Q)$ eine $\text{SPC}[\mathbf{S}]$ -Anfrage der Stelligkeit k .
- Sind Q und P zwei $\text{SPC}[\mathbf{S}]$ -Anfragen der Stelligkeiten m und n , so ist $(Q \times P)$ eine $\text{SPC}[\mathbf{S}]$ -Anfrage der Stelligkeit $m+n$.

Die *Semantik* $\llbracket Q \rrbracket$ von $\text{SPC}[\mathbf{S}]$ -Anfragen Q ist induktiv auf die offensichtliche Art definiert. Insbesondere für die Anfrage $Q := \{(c)\}$ ist die Semantik so definiert, dass für *jede* Datenbank \mathbf{I} gilt: $\llbracket Q \rrbracket(\mathbf{I}) = \{(c)\}$.

Folie 85

Beispiele für Anfragen der SPC-Algebra

- (4) Welche Kinos (Name & Adresse) spielen einen Film mit “Matt Damon”?

$$\pi_{7,8} \left(\sigma_{4=7} \left(\sigma_{1=5} \left(\sigma_{3 = \text{“Matt Damon”}} (\text{Filme} \times \text{Programm} \times \text{Kinos}) \right) \right) \right)$$

- (9) Egal, wie die DB aussieht, gib immer (“Terminator”, “Linda Hamilton”) aus!

$$\left(\{ \text{“Terminator”} \} \times \{ \text{“Linda Hamilton”} \} \right)$$

- (3) Gib Adresse und Telefonnummer des “Kino International” aus!

$$\pi_{2,4} \left(\sigma_{1=\text{“Kino International”}} (\text{Kinos}) \right)$$

- (5) Lläuft zur Zeit ein Film von “James Cameron”?

$$\pi \left(\sigma_{1=5} \left(\left(\sigma_{2=\text{“James Cameron”}} (\text{Filme}) \times \text{Programm} \right) \right) \right)$$

Der Operator $\pi(\cdot)$ bedeutet hier, dass auf „keine einzige“ Spalte projiziert wird. Das Ergebnis ist daher entweder $\{()\}$ oder \emptyset (also „ja“ oder „nein“).

- Sei **I** die Datenbank mit

$$\mathbf{I}(R) : \begin{array}{|c|c|} \hline 1 & 2 \\ \hline a & b \\ \hline a & c \\ \hline b & d \\ \hline \end{array} \quad \text{und} \quad \mathbf{I}(S) : \begin{array}{|c|c|c|} \hline 1 & 2 & 3 \\ \hline b & a & a \\ \hline c & b & a \\ \hline c & b & b \\ \hline \end{array}$$

und sei Q die Anfrage $\sigma_{2=5} \left((R \times \sigma_{2=3}(S)) \right)$.

Auswertung von Q in **I**:

- $\sigma_{2=3}(S)$ liefert auf **I** die Relation

1	2	3
b	a	a
c	b	b

- $(R \times \sigma_{2=3}(S))$ liefert auf **I** die Relations

1	2	3	4	5
a	b	b	a	a
a	b	c	b	b
a	c	b	a	a
a	c	c	b	b
b	d	b	a	a
b	d	c	b	b

- $\sigma_{2=5}((R \times \sigma_{2=3}(S)))$ liefert auf **I** die Relation

1	2	3	4	5
a	b	c	b	b

Somit ist $\llbracket Q \rrbracket(\mathbf{I}) = \{(a, b, c, b, b)\}$.

Folie 86

Verallgemeinerung des Selektions-Operators

Eine *positive konjunktive Selektionsbedingung* ist eine Formel F der Form

$$\gamma_1 \wedge \dots \wedge \gamma_n$$

wobei $n \geq 1$ und jedes γ_i eine Selektionsbedingung der Form $j_i=a_i$ oder $j_i=k_i$ für natürliche Zahlen $j_i, k_i \geq 1$ und Konstanten $a_i \in \mathbf{dom}$.

Der *Selektionsoperator* σ_F hat dieselbe Wirkung wie die Hintereinanderausführung der Selektionsoperatoren σ_{γ_i} für alle $i \in \{1, \dots, n\}$.

Folie 87

Eine Normalform für SPC-Anfragen

Sei **S** ein Relationsschema.

Definition 3.24. Eine SPC[**S**]-Anfrage ist in *Normalform*, falls sie von der Form

$$\pi_{j_1, \dots, j_k} \left(\{(c_1)\} \times \dots \times \{(c_m)\} \times \sigma_F(R_1 \times \dots \times R_\ell) \right)$$

ist, für $k, m, \ell \geq 0$, paarweise verschiedene Elemente j_1, \dots, j_k , so dass $\{1, \dots, m\} \subseteq \{j_1, \dots, j_k\}$, Konstanten $c_1, \dots, c_m \in \mathbf{dom}$, $R_1, \dots, R_\ell \in \mathbf{S}$ und F eine positive konjunktive Selektionsbedingung.

Proposition 3.25. *Für jede SPC[S]-Anfrage Q gibt es eine SPC[S]-Anfrage Q' in Normalform, die dieselbe Anfragefunktion definiert; und es gibt einen Polynomialzeit-Algorithmus, der Q' bei Eingabe von Q erzeugt.*

Beweis: Übung.

Folie 88

Benannte Perspektive: Die SPJR-Algebra

Operationen:

Selektion σ , Projektion π , Join \bowtie , Umbenennung (Renaming) δ

Attributnamen an Stelle von Spaltennummern!

Selektion: Zwei Varianten:

- *Operator $\sigma_{A=a}$,*
für eine Konstante $a \in \mathbf{dom}$ und einen *Attributnamen* A .
Dieser Operator kann angewendet werden auf R -Relationen I mit $A \in \mathit{sorte}(R)$ und liefert als Ausgabe die folgende Teilmenge der Relation I :

$$\sigma_{A=a}(I) := \{ t \in I \mid t(A) = a \}$$

- *Operator $\sigma_{A=B}$,* für zwei *Attributnamen* A und B .
Dieser Operator kann angewendet werden auf R -Relationen I mit $A, B \in \mathit{sorte}(R)$ und liefert als Ausgabe die folgende Teilmenge der Relation I :

$$\sigma_{A=B}(I) := \{ t \in I \mid t(A) = t(B) \}$$

„ $A=a$ “ und „ $A=B$ “ werden *Selektionsbedingungen* genannt.

Folie 89

Projektion:

Operator π_{A_1, \dots, A_k} , für paarweise verschiedene *Attributnamen* A_1, \dots, A_k (und $k \geq 0$).

Dieser Operator kann angewendet werden auf R -Relationen I mit $\mathit{sorte}(R) \supseteq \{A_1, \dots, A_k\}$ und liefert als Ausgabe die folgende Relation der Sorte $\{A_1, \dots, A_k\}$:

$$\pi_{A_1, \dots, A_k}(I) := \{ (A_1 : t(A_1), \dots, A_k : t(A_k)) \mid t \in I \}$$

Folie 90

An Stelle des Kartesischen Produkts: Natürlicher Join

Beispiel: Wenn I und J zwei Relationen mit *disjunkten Attributmengen* (d.h. Spaltenbezeichnungen) sind, so bilde einfach das herkömmliche Kartesische Produkt.

$$I \quad \bowtie \quad J \quad = \quad I \bowtie J$$

A	B
1	a
2	b
3	c

C	D	E
e	f	g
h	i	j

A	B	C	D	E
1	a	e	f	g
1	a	h	i	j
2	b	e	f	g
2	b	h	i	j
3	c	e	f	g
3	c	h	i	j

Frage: Was soll passieren, wenn I und J *gemeinsame Attribute* haben?

Antwort: Zueinander passende Tupel werden verschmolzen.

Beispiel:

$$I \quad \bowtie \quad J \quad = \quad I \bowtie J$$

A	B
1	a
2	b
3	c

B	C	D
a	f	g
a	i	j
c	l	m

A	B	C	D
1	a	f	g
1	a	i	j
3	c	l	m

Folie 91

Natürlicher Join:

Operator \bowtie

Dieser Operator kann angewendet werden auf eine R -Relation I und eine S -Relation J beliebiger Sorten und liefert als Ausgabe die folgende Relation der Sorte $\Sigma := \text{sorte}(R) \cup \text{sorte}(S)$:

$$I \bowtie J := \left\{ \text{Tupel } t \text{ der Sorte } \Sigma : \begin{array}{l} \text{es gibt Tupel } t' \in I \text{ und } t'' \in J \text{ so dass} \\ t|_{\text{sorte}(R)} = t' \text{ und } t|_{\text{sorte}(S)} = t'' \end{array} \right\}$$

Wir benutzen den Operator manchmal auch für einzelne Tupel:

Sind t' und t'' Tupel der Sorten $\text{sorte}(R)$ und $\text{sorte}(S)$, so schreiben wir $t' \bowtie t''$, um das Tupel t der Sorte $\text{sorte}(R) \cup \text{sorte}(S)$ mit $t|_{\text{sorte}(R)} = t'$ und $t|_{\text{sorte}(S)} = t''$ zu bezeichnen (bzw. den Wert „undefiniert“, falls es kein solches Tupel gibt, d.h. falls $t'|_{\text{sorte}(R) \cap \text{sorte}(S)} \neq t''|_{\text{sorte}(R) \cap \text{sorte}(S)}$).

Folie 92

Umbenennung (Renaming):

Operator δ_f , für eine Umbenennungsfunktion f , d.h. eine injektive Funktion $f : U \rightarrow \mathbf{att}$, für eine beliebige endliche Menge U von Attributnamen.

Dieser Operator kann angewendet werden auf R -Relationen I , für die gilt: $U \subseteq \text{sorte}(R)$ und $(\text{sorte}(R) \setminus U) \cap f(U) = \emptyset$ und liefert die folgende Relation der Sorte $f(U) \cup (\text{sorte}(R) \setminus U)$:

$$\delta_f(I) := \left\{ \begin{array}{l} \text{ex } t' \in I \text{ so dass gilt:} \\ \text{Tupel } t : \text{ f.a. } A \in U \text{ ist } t'(A) = t(f(A)) \\ \text{und f.a. } A \in \text{sorte}(R) \setminus U \text{ ist } t'(A) = t(A) \end{array} \right\}$$

Oft schreiben wir $A_1 \cdots A_k \mapsto B_1 \cdots B_k$ um die Umbenennungsfunktion f mit Definitionsbereich $U = \{A_1, \dots, A_k\}$ und Werten $f(A_i) = B_i$, für alle $i \in \{1, \dots, k\}$, zu bezeichnen.

Folie 93

Definition 3.26.

Sei \mathbf{S} ein Datenbankschema. Die Klasse der Anfragen der *SPJR-Algebra* über \mathbf{S} (kurz: $\text{SPJR}[\mathbf{S}]$) ist induktiv wie folgt definiert:

- Für alle Relationsnamen $R \in \mathbf{S}$ ist R eine $\text{SPJR}[\mathbf{S}]$ -Anfrage der Sorte $\text{sorte}(R)$.
- Für alle Konstanten $c \in \mathbf{dom}$ und alle Attributnamen $A \in \mathbf{att}$ ist $\{(A : c)\}$ eine $\text{SPJR}[\mathbf{S}]$ -Anfrage der Sorte $\{A\}$.
- Ist Q eine $\text{SPJR}[\mathbf{S}]$ -Anfrage der Sorte Σ , sind $A, B \in \Sigma$ und ist $a \in \mathbf{dom}$ eine Konstante, so sind auch $\sigma_{A=a}(Q)$ und $\sigma_{A=B}(Q)$ $\text{SPJR}[\mathbf{S}]$ -Anfragen der Sorte Σ .
- Ist Q eine $\text{SPJR}[\mathbf{S}]$ -Anfrage der Sorte Σ , ist $k \geq 0$ und sind A_1, \dots, A_k paarweise verschiedene Elemente aus Σ , so ist $\pi_{A_1, \dots, A_k}(Q)$ eine $\text{SPJR}[\mathbf{S}]$ -Anfrage der Sorte $\{A_1, \dots, A_k\}$.
- Sind Q und P zwei $\text{SPJR}[\mathbf{S}]$ -Anfragen der Sorten Σ und Π , so ist $(Q \bowtie P)$ eine $\text{SPJR}[\mathbf{S}]$ -Anfrage der Sorte $\Sigma \cup \Pi$.
- Ist Q eine $\text{SPJR}[\mathbf{S}]$ -Anfrage der Sorte Σ und ist $f : \Sigma \rightarrow \mathbf{att}$ eine Umbenennungsfunktion, so ist $\delta_f(Q)$ eine $\text{SPJR}[\mathbf{S}]$ -Anfrage der Sorte $f(\Sigma)$.

Die *Semantik* $\llbracket Q \rrbracket$ von $\text{SPJR}[\mathbf{S}]$ -Anfragen Q ist induktiv auf die offensichtliche Art definiert.

Insbesondere für die Anfrage $Q := \{(A : c)\}$ ist die Semantik so definiert, dass für jede Datenbank \mathbf{I} gilt: $\llbracket Q \rrbracket(\mathbf{I}) = \{(A : c)\}$.

Folie 94

Beispiele für Anfragen der SPJR-Algebra

- (4) Welche Kinos (Name & Adresse) spielen einen Film mit "Matt Damon"?

$$\pi_{Kino, Adresse} \left(\sigma_{\substack{\text{Schauspieler} = \\ \text{"Matt Damon"}}} (Filme \bowtie Programm \bowtie \delta_{\text{Name} \mapsto \text{Kino}} Kinos) \right)$$

- (9) Egal, wie die DB aussieht, gib immer ("Terminator", "Linda Hamilton") aus!

$$\left(\{ (1 : \text{"Terminator"}) \} \times \{ (2 : \text{"Linda Hamilton"}) \} \right)$$

- (3) Gib Adresse und Telefonnummer des "Kino International" aus!

$$\pi_{\text{Adresse, Telefon}} \left(\sigma_{\text{Name} = \text{"Kino International"}} (Kinos) \right)$$

- (5) Läuft zur Zeit ein Film von "James Cameron"?

$$\pi \left(\left(\sigma_{\text{Regie} = \text{"James Cameron"}} (Filme) \bowtie Programm \right) \right)$$

Der Operator $\pi(\cdot)$ bedeutet hier, dass auf „keine einzige“ Spalte projiziert wird. Das Ergebnis ist daher entweder $\{()\}$ oder \emptyset (also „ja“ oder „nein“).

- Sei **I** die Datenbank mit

$$\mathbf{I}(R) : \begin{array}{|c|c|} \hline A & B \\ \hline a & b \\ \hline a & c \\ \hline b & d \\ \hline \end{array} \quad \text{und} \quad \mathbf{I}(S) : \begin{array}{|c|c|c|} \hline C & D & E \\ \hline b & a & a \\ \hline c & b & a \\ \hline c & b & b \\ \hline \end{array}$$

und sei Q die Anfrage $\sigma_{C=c} \left(\pi_{A,C} \left((R \bowtie \delta_{E \mapsto A}(S)) \right) \right)$.

Auswertung von Q in **I**:

- $\delta_{E \mapsto A}(S)$ liefert auf **I** die Relation

C	D	A
b	a	a
c	b	a
c	b	b

- $(R \bowtie \delta_{E \rightarrow A}(S))$ liefert auf **I** die Relations

A	B	C	D
a	b	b	a
a	b	c	b
a	c	b	a
a	c	c	b
b	d	c	b

- $\pi_{A,C}((R \bowtie \delta_{E \rightarrow A}(S)))$ liefert auf **I** die Relation

A	C
a	b
a	c
a	b
a	c
b	c

=

A	C
a	b
a	c
b	c

- $\sigma_{C=c}(\pi_{A,C}((R \bowtie \delta_{E \rightarrow A}(S))))$ liefert auf **I** die Relation

A	C
a	c
b	c

Somit ist $\llbracket Q \rrbracket(\mathbf{I}) = \{(A : a, C : c), (A : b, C : c)\}$.

Folie 95

Verallgemeinerung des Selektions-Operators

Wie bei der SPC-Algebra lassen wir wieder eine *Verallgemeinerung des Selektions-Operators* zu:

- Eine *positive konjunktive Selektionsbedingung* ist eine Formel F der Form

$$\gamma_1 \wedge \dots \wedge \gamma_n$$

wobei $n \geq 1$ und jedes γ_i eine Selektionsbedingung der Form $A_i = a_i$ oder $A_i = B_i$ für Attributnamen A_i, B_i und Konstanten $a_i \in \mathbf{dom}$.

- Der *Selektionsoperator* σ_F hat dieselbe Wirkung wie die Hintereinanderausführung der Selektionsoperatoren σ_{γ_i} für alle $i \in \{1, \dots, n\}$.

Folie 96

Eine Normalform für SPJR-Anfragen

Sei \mathbf{S} ein Relationenschema.

Definition 3.27.

Eine SPJR[\mathbf{S}]-Anfrage ist in *Normalform*, falls sie von der Form

$$\pi_{B_1, \dots, B_k} \left(\{(A_1 : c_1)\} \bowtie \dots \bowtie \{(A_m : c_m)\} \bowtie \sigma_F(\delta_{f_1}(R_1) \bowtie \dots \bowtie \delta_{f_\ell}(R_\ell)) \right)$$

ist, für $k, m, \ell \geq 0$, $B_1, \dots, B_k, A_1, \dots, A_m \in \mathbf{att}$ so dass $\{A_1, \dots, A_m\} \subseteq \{B_1, \dots, B_k\}$ und die A_1, \dots, A_m paarweise verschieden, $c_1, \dots, c_m \in \mathbf{dom}$, $R_1, \dots, R_\ell \in \mathbf{S}$, eine positive konjunktive Selektionsbedingung F und Umbenennungsfunktionen f_1, \dots, f_ℓ , so dass die Sorten von $\delta_{f_1}(R_1), \dots, \delta_{f_\ell}(R_\ell)$ paarweise disjunkt sind und keins der A_1, \dots, A_m als Attribut von einem der $\delta_{f_j}(R_j)$ vorkommt.

Proposition 3.28. *Für jede SPJR[\mathbf{S}]-Anfrage Q gibt es eine SPJR[\mathbf{S}]-Anfrage Q' in Normalform, die dieselbe Anfragefunktion definiert; und es gibt einen Polynomialzeit-Algorithmus, der Q' bei Eingabe von Q erzeugt.*

Beweis: Übung.

Folie 97

Nicht-Erfüllbare Anfragen

Bemerkung:

Sowohl in der SPC-Algebra als auch in der SPJR-Algebra lassen sich unerfüllbare Anfragen ausdrücken.

Beispiel: Die Anfrage $Q :=$

$$\sigma_{3="George Clooney"} \left(\sigma_{3="Matt Damon"}(Filme) \right)$$

wählt in einer Datenbank vom Schema **KINO** genau diejenigen Tupel $t = (a, b, c)$ aus der *Filme*-Relation aus, für deren dritte Komponente c gilt: $c = \text{"Matt Damon"}$ und $c = \text{"George Clooney"}$.

Solche Tupel kann es aber nicht geben!

Für jede Datenbank \mathbf{I} vom Schema **KINO** gilt also: $\llbracket Q \rrbracket(\mathbf{I}) = \emptyset$.

Somit ist die Anfrage Q nicht erfüllbar.

Folie 98

Äquivalenz der Ausdruckstärke der SPC-Algebra und der SPJR-Algebra

Lemma 3.29. *Die SPC-Algebra und die SPJR-Algebra können genau dieselben Anfragen ausdrücken.*

Es gilt sogar: Jede Anfrage aus einer dieser Anfragesprachen kann in polynomialer Zeit in eine äquivalente Anfrage der anderen Anfragesprache übersetzt werden.

Beweis. Wir nutzen die Normalformen für SPC- und für SPJR-Anfragen, die in den Propositionen 3.25 und 3.28 bereitgestellt wurden. Es reicht daher o.B.d.A. aus, jeweils nur solche Anfragen zu übersetzen, die in Normalform sind.

SPC \rightsquigarrow *SPJR*:

Bei Eingabe einer SPC-Anfrage Q der Form

$$\pi_{j_1, \dots, j_k} \left(\{(c_1)\} \times \dots \times \{(c_m)\} \times \sigma_F(R_1 \times \dots \times R_\ell) \right)$$

konstruieren wir die SPJR-Anfrage Q' der Form

$$\pi_{j_1, \dots, j_k} \left(\{(1 : c_1)\} \bowtie \dots \bowtie \{(m : c_m)\} \bowtie \sigma_{F'}(\delta_{f_1}(R_1) \bowtie \dots \bowtie \delta_{f_\ell}(R_\ell)) \right),$$

wobei f_1, \dots, f_ℓ Umbenennungsfunktionen sind, die Folgendes gewährleisten:

- $\delta_{f_1}(R_1)$ hat die Spaltenbezeichner $m+1, \dots, m+\text{ar}(R_1)$,
- $\delta_{f_2}(R_2)$ hat die Spaltenbezeichner $m+\text{ar}(R_1)+1, \dots, m+\text{ar}(R_1)+\text{ar}(R_2)$,
- \vdots
- $\delta_{f_\ell}(R_\ell)$ hat die Spaltenbezeichner $m_{\ell-1}+1, \dots, m_{\ell-1} + \text{ar}(R_\ell)$, wobei $m_{\ell-1} := m + \text{ar}(R_1) + \dots + \text{ar}(R_{\ell-1})$ ist.

Die verallgemeinerte Selektionsbedingung F' entsteht dabei aus der verallgemeinerten Selektionsbedingung F , indem jede Selektionsbedingung in F der Form

- $\nu_i = a_i$ (für eine natürliche Zahl ν_i und eine Konstante $a_i \in \mathbf{dom}$) ersetzt wird durch die Selektionsbedingung $\nu'_i = a_i$, für $\nu'_i := m + \nu_i$, und
- $\nu_i = \mu_i$ (für natürliche Zahlen ν_i und μ_i) ersetzt wird durch die Selektionsbedingung $\nu'_i = \mu'_i$, für $\nu'_i := m + \nu_i$ und $\mu'_i := m + \mu_i$.

Einfaches Nachprüfen zeigt, dass die beiden Anfragen Q und Q' tatsächlich äquivalent sind, d.h. dass für jede Datenbank \mathbf{I} vom zu Q passenden Schema gilt: $\llbracket Q \rrbracket(\mathbf{I}) = \llbracket Q' \rrbracket(\mathbf{I})$. Details: Übung!

$SPJR \rightsquigarrow SPC$:

Analog; Details: Übung!

□

Folie 99

Äquivalenz des deskriptiven und des algebraischen Ansatzes

Theorem 3.30 (Äquivalenz konjunktiver Anfragesprachen).

Die folgenden Anfragesprachen können genau dieselben erfüllbaren Anfragefunktionen ausdrücken:

- die Klasse der regelbasierten konjunktiven Anfragen
- die Klasse der Tableau-Anfragen
- die Klasse der Anfragen des konjunktiven Kalküls
- die Anfragen der SPC-Algebra
- die Anfragen der SPJR-Algebra.

Es gilt sogar: Jede Anfrage aus einer dieser Anfragesprachen kann in polynomieller Zeit in äquivalente Anfragen der anderen Anfragesprachen übersetzt werden.

Beweis. Von Lemma 3.14 wissen wir bereits, dass die Anfragesprachen aus (a), (b) und (c) dieselbe Ausdrucksstärke besitzen, und dass Anfragen jeweils in Polynomialzeit in jede der drei Sprachen übersetzt werden können.

Von Lemma 3.29 wissen wir, dass SPC-Algebra und SPJR-Algebra dieselbe Ausdrucksstärke besitzen, und dass Anfragen in Polynomialzeit in äquivalente Anfragen der beiden Algebren übersetzt werden können.

Die Einschränkung auf *erfüllbare* Anfragen müssen wir machen, weil gemäß Satz 3.6 alle regelbasierten konjunktiven Anfragen erfüllbar sind, während in der SPC- und der SPJR-Algebra auch nicht-erfüllbare Anfragen formuliert werden können.

Um den Beweis von Theorem 3.30 abzuschließen, genügt es, die Richtung von (a) nach (d) und die Richtung von (d) nach (a) zu beweisen.

(a) \rightsquigarrow (d):

Sei Q eine regelbasierte konjunktive Anfrage der Form

$$\text{Ans}(u) \leftarrow R_1(u_1), \dots, R_\ell(u_\ell).$$

Sei $r := \text{ar}(\text{Ans})$, und für jedes $i \in \{1, \dots, \ell\}$ sei $r_i := \text{ar}(R_i)$.

Sei u von der Form x_1, \dots, x_r , sei m die Anzahl der Konstanten in x_1, \dots, x_r und seien $\alpha_1 < \dots < \alpha_m$ diejenigen Indizes in $\{1, \dots, r\}$ mit $x_{\alpha_1}, \dots, x_{\alpha_m} \in \mathbf{dom}$.

Außerdem sei u_1 von der Form y_1, \dots, y_{r_1} , sei u_2 von der Form $y_{r_1+1}, \dots, y_{r_1+r_2}$, usw., und sei u_ℓ von der Form $y_{m_{\ell-1}+1}, \dots, y_{m_{\ell-1}+r_\ell}$, wobei $m_{\ell-1} := r_1 + \dots + r_{\ell-1}$ ist und jeweils $y_i \in \mathbf{var} \cup \mathbf{dom}$ für jedes $i \in \{1, \dots, m_{\ell-1}+r_\ell\}$ gilt.

Wir wählen folgende SPC-Algebra-Anfrage Q' :

$$\pi_{j_1, \dots, j_r} \left(\{(x_{\alpha_1})\} \times \dots \times \{(x_{\alpha_m})\} \times \sigma_F(R_1 \times \dots \times R_\ell) \right),$$

wobei die verallgemeinerte Selektionsbedingung F und die Indizes j_1, \dots, j_r wie folgt gewählt sind. F ist die Konjunktion aller Bedingungen der Form

- $i = c$, für alle $i \in \{1, \dots, m_{\ell-1}+r_\ell\}$ und alle $c \in \mathbf{dom}$ mit $y_i = c$, und
- $i = j$, für alle $i, j \in \{1, \dots, m_{\ell-1}+r_\ell\}$ mit $i \neq j$ und $y_i = y_j$.

Die Indizes j_1, \dots, j_r sind so gewählt, dass für jedes $i \in \{1, \dots, r\}$ gilt:

$$j_i := \begin{cases} m+k & , \text{ falls } x_i \in \mathbf{var} \text{ und } k \text{ minimal, so dass } x_i = y_k \\ k & , \text{ falls } x_i \in \mathbf{dom} \text{ und } k \in \{1, \dots, m\}, \text{ so dass } i = \alpha_k \end{cases}$$

Einfaches Nachprüfen zeigt, dass Q und Q' äquivalent sind. Details: Übung!

(d) \rightsquigarrow (a):

Wir können hier rekursiv entlang dem Aufbau von Anfragen der SPC-Algebra vorgehen (und beachten dabei, dass bei einer *erfüllbaren* Anfrage Q der SPC-Algebra auch sämtliche Teilanfragen \hat{Q} , aus denen Q aufgebaut ist, erfüllbar sind). Details: Übung! □

Beispiel zur Übersetzung von Anfragen

Betrachte die Anfrage $Q :=$

$$\begin{aligned} \text{Ans}(x, \text{“Treffer”}) \leftarrow & \text{Programm}(x, y, z_1), \\ & \text{Filme}(z_2, y, \text{“George Clooney”}), \\ & \text{Filme}(z_3, \text{“George Clooney”}, y) \end{aligned}$$

Unser Beweis der Richtung (a) \rightsquigarrow (b) von Theorem 3.30 übersetzt diese Anfrage in die SPC-Algebra-Anfrage

$$\pi_{2,1} \left(\{(\text{“Treffer”})\} \times \sigma_F(\text{Programm} \times \text{Filme} \times \text{Filme}) \right)$$

mit der verallgemeinerten Selektionsbedingung $F :=$

$$6 = \text{“George Clooney”} \wedge 8 = \text{“George Clooney”} \wedge 2 = 5 \wedge 2 = 9 \wedge 5 = 9.$$

3.4 Homomorphismus-Satz, Statische Analyse und Anfrageminimierung

Folie 101

Vorbemerkung zum Thema „Statische Analyse, Optimierung“

Optimierung:

Finde zur gegebenen Anfrage eine „minimale“ äquivalente Anfrage

Definition 3.31 (Äquivalenz und Query Containment).

Seien P und Q Anfragen einer Anfragesprache über einem DB-Schema \mathbf{S} .

- (a) Wir schreiben $Q \equiv P$ und sagen „ Q ist äquivalent zu P “, falls für alle Datenbanken $\mathbf{I} \in \text{inst}(\mathbf{S})$ gilt: $\llbracket Q \rrbracket(\mathbf{I}) = \llbracket P \rrbracket(\mathbf{I})$.
- (b) Wir schreiben $Q \sqsubseteq P$ und sagen „ Q ist in P enthalten“, falls für alle Datenbanken $\mathbf{I} \in \text{inst}(\mathbf{S})$ gilt: $\llbracket Q \rrbracket(\mathbf{I}) \subseteq \llbracket P \rrbracket(\mathbf{I})$.

Statische Analyse:

- *Äquivalenzproblem:*
Bei Eingabe zweier Anfragen Q, P teste, ob $Q \equiv P$.
- *Query Containment Problem:*
Bei Eingabe von Anfragen Q, P teste, ob $Q \sqsubseteq P$.

- *Erfüllbarkeitsproblem:*
Bei Eingabe einer Anfrage Q teste, ob Q erfüllbar ist (d.h. ob es eine DB \mathbf{I} gibt, so dass $\llbracket Q \rrbracket(\mathbf{I}) \neq \emptyset$ ist).

Bekannt:

- Für regelbasierte konjunktive Anfragen ist das Erfüllbarkeitsproblem trivial (Satz 3.6).
- Für konjunktive Anfragen mit „=“ (bzw. für SPC- bzw. SPJR-Anfragen) ist das Erfüllbarkeitsproblem in polynomieller Zeit lösbar (dies wurde in den Übungsaufgaben gezeigt).
- *Jetzt:* Algorithmen für's Query Containment Problem und für's Äquivalenzproblem für konjunktive Anfragen.

Folie 102

Zusammenhänge:

- *Äquivalenz vs. Containment:*
 - $Q \equiv P \iff Q \sqsubseteq P \text{ und } P \sqsubseteq Q$
 - $Q \sqsubseteq P \iff (Q \vee P) \equiv P$
Dies lässt sich für Optimierung nutzen!
- *Containment vs. Erfüllbarkeit:*
 - Q unerfüllbar $\implies Q \sqsubseteq P$ für alle Anfragen P
 - $Q \sqsubseteq P \iff (Q \wedge \neg P)$ ist unerfüllbar
Dies lässt sich für Optimierung nutzen!
- *Erfüllbarkeit vs. Äquivalenz:*
 - Q unerfüllbar $\implies (Q \vee P) \equiv P$ für alle Anfragen P
Dies lässt sich für Optimierung nutzen!

Folie 103

Zur Erinnerung: Tableau-Anfragen — Beispiel

Beispiel-Anfrage:

Filmtitel + Regisseur aller z.Zt. laufenden Filme, deren Regisseur schon mal mit “Sandra Bullock” zusammengearbeitet hat.

Als regelbasierte konjunktive Anfrage:

$$\begin{aligned} \text{Ans}(x_T, x_R) \leftarrow & \text{Programm}(x_K, x_T, x_Z), \\ & \text{Filme}(x_T, x_R, x_S), \\ & \text{Filme}(y_T, x_R, \text{“Sandra Bullock”}) \end{aligned}$$

Als Tableau-Anfrage: $(\mathbb{T}, (x_T, x_R))$ mit folgendem Tableau \mathbb{T} :

<i>Programm</i>	Kino	Titel	Zeit
	x_K	x_T	x_Z
<i>Filme</i>	Titel	Regie	Schauspieler
	x_T	x_R	x_S
	y_T	x_R	“Sandra Bullock”

Folie 104

Homomorphismen

Definition 3.32. Seien $Q' = (\mathbb{T}', u')$ und $Q = (\mathbb{T}, u)$ zwei Tableau-Anfragen über einem DB-Schema \mathbf{S} .

(a) Eine *Substitution für Q'* ist eine Abbildung $h : \text{var}(Q') \rightarrow \mathbf{var} \cup \mathbf{dom}$.

Wie üblich setzen wir h auf natürliche Weise fort zu einer Abbildung von $\text{var}(Q') \cup \mathbf{dom}$ nach $\mathbf{var} \cup \mathbf{dom}$, so dass $h|_{\mathbf{dom}} = \text{id}$.

Für ein freies Tupel $t = (e_1, \dots, e_k)$ setzen wir $h(t) := (h(e_1), \dots, h(e_k))$.

Für eine Menge M von freien Tupeln ist $h(M) := \{h(t) : t \in M\}$.

(b) Eine Substitution h für Q' heißt *Homomorphismus von Q' auf Q* , falls

- $h(u') = u$ und
- $h(\mathbb{T}') \subseteq \mathbb{T}$, d.h. für alle $R \in \mathbf{S}$ ist $h(\mathbb{T}'(R)) \subseteq \mathbb{T}(R)$.

Beachte: Dann gilt insbes.: $h(\text{var}(Q')) \subseteq \text{var}(Q) \cup \text{adom}(Q)$.

Beispiel: $\mathbf{S} := \{R\}$, $Q' := (\mathbb{T}', (x, y))$, $Q := (\mathbb{T}, (x, y))$ mit

$$\mathbb{T}'(R) := \begin{array}{|c|c|} \hline A & B \\ \hline x & y_1 \\ \hline x_1 & y_1 \\ \hline x_1 & y \\ \hline \end{array} \quad \mathbb{T}(R) := \begin{array}{|c|c|} \hline A & B \\ \hline x & y \\ \hline \end{array}$$

Homomorphismus h von Q' auf Q : $h : x, y, x_1, y_1 \mapsto x, y, x, y$.

Es gibt keinen Homomorphismus von Q auf Q' .

Folie 105

Die kanonische Datenbank $\mathbf{I}_Q^{Q'}$

Idee: $\mathbf{I}_Q^{Q'}$ ist die Datenbank, die aus dem Tableau \mathbb{T} von Q entsteht, indem jede Variable in \mathbb{T} durch eine Konstante repräsentiert wird, die nicht in Q oder Q' vorkommt.

Wir legen ein für alle Mal für jedes endliche $C \subseteq \mathbf{dom}$ eine *injektive* Abbildung $\alpha_C : \mathbf{var} \rightarrow \mathbf{dom} \setminus C$ fest. Wie üblich setzen wir α_C fort zu einer Abbildung von $\mathbf{var} \cup \mathbf{dom}$ nach \mathbf{dom} mit $\alpha|_{\mathbf{dom}} = \text{id}$.

Für die folgendermaßen definierte „Umkehrfunktion“

$$\alpha_C^{-1} : \mathbf{dom} \rightarrow \mathbf{var} \cup \mathbf{dom}$$

$$\alpha_C^{-1}(a) := \begin{cases} a & \text{falls } a \notin \text{Bild}(\alpha_C) \\ y & \text{falls } a = \alpha_C(y) \text{ für } y \in \mathbf{var} \end{cases}$$

gilt für alle $b \in \mathbf{var} \cup C$, dass $\alpha_C^{-1}(\alpha_C(b)) = b$.

Definition 3.33 (Idee: Repräsentiere $Q = (\mathbb{T}, u)$ durch eine Datenbank $\mathbf{I}_Q^{Q'}$ und ein Tupel $u_Q^{Q'}$).

$Q' = (\mathbb{T}', u')$ und $Q = (\mathbb{T}, u)$ seien Tableau-Anfragen über einem DB-Schema \mathbf{S} . Die *kanonische Datenbank* $\mathbf{I}_Q^{Q'} \in \text{inst}(\mathbf{S})$ und das

kanonische Tupel $u_Q^{Q'}$ sind folgendermaßen definiert:

Für $C := \text{adom}(Q) \cup \text{adom}(Q')$ ist $u_Q^{Q'} := \alpha_C(u)$ und $\mathbf{I}_Q^{Q'} := \alpha_C(\mathbb{T})$, d.h. $\mathbf{I}_Q^{Q'}(R) = \alpha_C(\mathbb{T}(R))$, für alle $R \in \mathbf{S}$.

Proposition 3.34. $Q' = (\mathbb{T}', u')$ und $Q = (\mathbb{T}, u)$ seien Tableau-Anfragen über einem DB-Schema \mathbf{S} . Dann gilt:

Es gibt einen Homomorphismus von Q' auf $Q \iff u_Q^{Q'} \in \llbracket Q' \rrbracket(\mathbf{I}_Q^{Q'})$.

Beweis. Sei $C := \text{adom}(Q) \cup \text{adom}(Q')$.

„ \implies “: Sei $h : \text{Var}(Q') \rightarrow \mathbf{var} \cup \mathbf{dom}$ ein Homomorphismus von Q' auf Q .

D.h.: $h(u') = u$ und $h(\mathbb{T}') \subseteq \mathbb{T}$.

Unser Ziel ist, zu zeigen, dass $u_Q^{Q'} \in \llbracket Q' \rrbracket(\mathbf{I}_Q^{Q'})$. Dazu müssen wir eine Belegung $\beta : \text{Var}(Q') \rightarrow \mathbf{dom}$ finden, für die gilt:

$$(1) \quad \beta(u') = u_Q^{Q'} \quad \text{und}$$

$$(2) \quad \beta \text{ ist eine Einbettung von } Q' \text{ in } \mathbf{I}_Q^{Q'}, \text{ d.h. es gilt: } \beta(\mathbb{T}') \subseteq \mathbf{I}_Q^{Q'}.$$

Dazu betrachten wir die Belegung $\beta : \text{Var}(Q') \rightarrow \mathbf{dom}$ mit $\beta := \alpha_C \circ h$, d.h. $\beta(x) = \alpha_C(h(x))$ für alle $x \in \text{Var}(Q')$.

Offensichtlicherweise gilt dann:

$$(1) \quad \beta(u') = \alpha_C(h(u')) = \alpha_C(u) \stackrel{\text{Def.}}{=} u_Q^{Q'}.$$

$$(2) \quad \beta(\mathbb{T}') = \alpha_C(h(\mathbb{T}')) \stackrel{\text{Hom.}}{\subseteq} \alpha_C(\mathbb{T}) \stackrel{\text{Def.}}{=} \mathbf{I}_Q^{Q'}.$$

Also ist $u_Q^{Q'} \in \llbracket Q' \rrbracket(\mathbf{I}_Q^{Q'})$.

„ \impliedby “: Laut Voraussetzung gilt $u_Q^{Q'} \in \llbracket Q' \rrbracket(\mathbf{I}_Q^{Q'})$. Es gibt also eine Einbettung $\beta : \text{Var}(Q') \rightarrow \mathbf{dom}$ von Q' in $\mathbf{I}_Q^{Q'}$ mit $\beta(u') = u_Q^{Q'}$.

Unser Ziel ist, einen Homomorphismus h von Q' auf Q zu finden.

Betrachte dazu die Substitution $h : \text{Var}(Q') \rightarrow \mathbf{var} \cup \mathbf{dom}$ mit

$h := \alpha_C^{-1} \circ \beta$, d.h. $\alpha_C(h(x)) = \beta(x)$ gilt für alle $x \in \text{Var}(Q')$.

Um nachzuweisen, dass h ein Homomorphismus von Q' auf Q ist, müssen wir zeigen, dass Folgendes gilt:

$$(1) \quad h(u') = u \quad \text{und}$$

$$(2) \quad h(\mathbb{T}') \subseteq \mathbb{T}.$$

$$\text{zu (1): } h(u') = \alpha_C^{-1}(\beta(u')) = \alpha_C^{-1}(u_Q^{Q'}) = u.$$

$$\text{zu (2): } h(\mathbb{T}') = \alpha_C^{-1}(\beta(\mathbb{T}')). \text{ Da } \beta \text{ eine Einbettung von } Q' \text{ in } \mathbf{I}_Q^{Q'} \text{ ist, gilt: } \beta(\mathbb{T}') \subseteq \mathbf{I}_Q^{Q'}. \text{ Somit ist } h(\mathbb{T}') = \alpha_C^{-1}(\beta(\mathbb{T}')) \subseteq \alpha_C^{-1}(\mathbf{I}_Q^{Q'}) = \mathbb{T}.$$

Also ist h ein Homomorphismus von Q' auf Q . □

Beispiel:

Sei $\mathbf{S} := \{R\}$, $Q' := (\mathbb{T}', (x, y))$ und $Q := (\mathbb{T}, (x, y))$ mit

$$\mathbb{T}'(R) = \begin{array}{|c|c|} \hline A & B \\ \hline x & y_1 \\ \hline x_1 & y_1 \\ \hline x_1 & y \\ \hline \end{array} \qquad \mathbb{T}(R) = \begin{array}{|c|c|} \hline A & B \\ \hline x & y \\ \hline y & d \\ \hline \end{array}$$

wobei $d \in \mathbf{dom}$ ist.

Um die kanonischen Datenbanken und die kanonischen Tupel zu Q und Q' zu bilden, betrachten wir die Menge $C := \text{adom}(Q) \cup \text{adom}(Q') = \{d\}$. Für jedes $z \in \mathbf{var}$ sei $a_z := \alpha_C(z)$. D.h. a_z ist die Konstante, die gemäß α_C die Variable z repräsentiert. Dann ist $d \notin \{a_z : z \in \mathbf{var}\}$ und es gilt:

Kanonische Tupel: $u_{Q'}^Q = (a_x, a_y)$ und $u_Q^{Q'} = (a_x, a_y)$.

Kanonische Datenbanken:

$$\mathbf{I}_{Q'}^Q(R) = \begin{array}{|c|c|} \hline A & B \\ \hline a_x & a_{y_1} \\ \hline a_{x_1} & a_{y_1} \\ \hline a_{x_1} & a_y \\ \hline \end{array} \qquad \mathbf{I}_Q^{Q'}(R) = \begin{array}{|c|c|} \hline A & B \\ \hline a_x & a_y \\ \hline a_y & d \\ \hline \end{array}$$

Um herauszufinden, ob $Q \sqsubseteq Q'$ ist, genügt es laut Theorem 3.35 zu testen, ob $u_{Q'}^Q \in \llbracket Q' \rrbracket(\mathbf{I}_{Q'}^Q)$ ist.

Genauer Betrachten der Anfrage Q' und der Datenbank $\mathbf{I}_{Q'}^Q$ zeigt, dass die Abbildung β mit $\beta(x) := \beta(x_1) := a_x$ und $\beta(y) := \beta(y_1) := a_y$ eine Einbettung von \mathbb{T}' in $\mathbf{I}_{Q'}^Q$ ist, für die gilt: $\beta((x, y)) = (a_x, a_y) = u_{Q'}^Q$. Somit ist also $u_{Q'}^Q \in \llbracket Q' \rrbracket(\mathbf{I}_{Q'}^Q)$. Gemäß Theorem 3.35 gilt also: $Q \sqsubseteq Q'$.

Um herauszufinden, ob $Q' \sqsubseteq Q$ ist, genügt es laut Theorem 3.35 zu testen, ob $u_Q^{Q'} \in \llbracket Q \rrbracket(\mathbf{I}_Q^{Q'})$ ist.

Genauer Betrachten der Anfrage Q und der Datenbank $\mathbf{I}_Q^{Q'}$ zeigt, dass es keine Einbettung von \mathbb{T} in $\mathbf{I}_Q^{Q'}$ gibt (denn angenommen, β wäre eine solche Einbettung, dann müsste das Tupel $(y, d) \in \mathbb{T}(R)$ durch β auf ein Tupel in $\mathbf{I}_Q^{Q'}(R)$ abgebildet werden — aber $\beta((y, d)) = (\beta(y), d)$, und in $\mathbf{I}_Q^{Q'}(R)$ gibt es kein Tupel, das in der zweiten Komponente den Eintrag d hat).

Somit ist also $\llbracket Q \rrbracket(\mathbf{I}_Q^{Q'}) = \emptyset$. Daher ist $u_Q^{Q'} \notin \llbracket Q \rrbracket(\mathbf{I}_Q^{Q'})$. Gemäß Theorem 3.35 gilt also: $Q' \not\sqsubseteq Q$.

Der Homomorphismus-Satz

Theorem 3.35 (Chandra, Merlin, 1977).

Sei \mathbf{S} ein Datenbankschema und seien $Q' = (\mathsf{T}', u')$ und $Q = (\mathsf{T}, u)$ zwei Tableau-Anfragen über \mathbf{S} . Dann gilt:

$$\begin{aligned} Q &\sqsubseteq Q' \\ \iff &\text{ es gibt einen Homomorphismus von } Q' \text{ auf } Q \\ \iff &u_Q^{Q'} \in \llbracket Q' \rrbracket(\mathbf{I}_Q^{Q'}). \end{aligned}$$

Beweis.

Die Äquivalenz der letzten beiden Zeilen gilt gemäß Proposition 3.34.

Für die Richtung von der zweiten zur ersten Zeile gibt es laut

Voraussetzung einen Homomorphismus h von Q' auf Q .

Sei \mathbf{I} eine beliebige Datenbank vom Schema \mathbf{S} . Wir müssen zeigen, dass $\llbracket Q \rrbracket(\mathbf{I}) \subseteq \llbracket Q' \rrbracket(\mathbf{I})$ ist.

Sei dazu t ein beliebiges Tupel in $\llbracket Q \rrbracket(\mathbf{I})$. Wir müssen zeigen, dass

$t \in \llbracket Q' \rrbracket(\mathbf{I})$ ist. Gemäß der Definition der Semantik von Tableau-Anfragen

reicht es, eine Einbettung β' von T' in \mathbf{I} finden, so dass $t = \beta'(u')$ ist.

Laut Voraussetzung ist $t \in \llbracket Q \rrbracket(\mathbf{I})$. Es gibt also eine Einbettung β von T in \mathbf{I} , so dass $t = \beta(u)$ ist.

Außerdem gibt es laut Voraussetzung einen Homomorphismus h von Q' auf Q , d.h. eine Abbildung $h : \text{Var}(Q') \rightarrow \mathbf{var} \cup \mathbf{dom}$, so dass $h(u') = u$ und $h(\mathsf{T}') \subseteq \mathsf{T}$ ist.

Wir wählen $\beta' := \beta \circ h$. D.h. β' ist die Abbildung von $\text{Var}(Q')$ nach \mathbf{dom} mit $\beta'(x) := \beta(h(x))$ für alle $x \in \text{Var}(Q')$. Für dieses β' gilt:

- $\beta'(u') = \beta(h(u')) \stackrel{\text{Hom.}}{=} \beta(u) = t$ und
- $\beta'(\mathsf{T}') = \beta(h(\mathsf{T}')) \stackrel{\text{Hom.}}{\subseteq} \beta(\mathsf{T}) \subseteq \mathbf{I}$.

Somit ist β' eine Einbettung von T' in \mathbf{I} mit $\beta'(u') = t$. Daher ist $t \in \llbracket Q' \rrbracket(\mathbf{I})$.

Für die Richtung von der ersten Zeile zur dritten Zeile gilt laut

Voraussetzung, dass $Q \sqsubseteq Q'$. Unser Ziel ist, zu zeigen, dass für $\mathbf{I} := \mathbf{I}_Q^{Q'}$ gilt:

$$u_Q^{Q'} \in \llbracket Q' \rrbracket(\mathbf{I}).$$

Wegen $Q \sqsubseteq Q'$ gilt insbesondere für die Datenbank $\mathbf{I} = \mathbf{I}_Q^{Q'}$, dass

$\llbracket Q \rrbracket(\mathbf{I}) \subseteq \llbracket Q' \rrbracket(\mathbf{I})$. Daher genügt es, zu zeigen, dass $u_Q^{Q'} \in \llbracket Q \rrbracket(\mathbf{I})$ ist.

Sei β die Belegung mit $\beta(x) := \alpha_C(x)$ für alle $x \in \text{Var}(Q)$. Dann gilt offensichtlich:

- $\beta(u) = \alpha_C(u) \stackrel{\text{Def.}}{=} u_Q^{Q'}$ und
- $\beta(T) = \alpha_C(T) \stackrel{\text{Def.}}{=} \mathbf{I}_Q^{Q'}$.

Also ist β eine Einbettung von T in $\mathbf{I}_Q^{Q'}$ mit $\beta(u) = u_Q^{Q'}$. Gemäß der Definition der Semantik von Tableau-Anfragen ist also $u_Q^{Q'} \in \llbracket Q \rrbracket(\mathbf{I}_Q^{Q'})$. \square

Korollar 3.36. *Das*

QUERY CONTAINMENT PROBLEM FÜR TABLEAU-ANFRAGEN

Eingabe: Tableau-Anfragen Q und Q' über einem DB-Schema S

Frage: Ist $Q \sqsubseteq Q'$
(d.h. gilt für alle $\mathbf{I} \in \text{inst}(S)$, dass $\llbracket Q \rrbracket(\mathbf{I}) \subseteq \llbracket Q' \rrbracket(\mathbf{I})$?

ist NP-vollständig.

Beweis.

Zugehörigkeit zu NP:

Bei Eingabe von Q und Q' (und der Frage, ob $Q \sqsubseteq Q'$ ist), gehen wir wie folgt vor:

- (1) Konstruiere die kanonische Datenbank $\mathbf{I}_Q^{Q'}$ und das kanonische Tupel $u_Q^{Q'}$.
Das geht deterministisch in Zeit polynomiell in der Größe von Q und Q' .
- (2) Löse das Auswertungsproblem für die Datenbank $\mathbf{I}_Q^{Q'}$ und die Boolesche konjunktive Anfrage

„Liegt $u_Q^{Q'}$ in $\llbracket Q' \rrbracket(\mathbf{I}_Q^{Q'})$?“

Dies geht nichtdeterministisch in polynomieller Zeit (durch „Raten“ einer Belegung β und anschließendes Verifizieren, dass $\beta(u') = u_Q^{Q'}$ und $\beta(T') \subseteq \mathbf{I}_Q^{Q'}$ gilt).

NP-Härte:

Wir wissen bereits, dass das Auswertungsproblem für Boolesche Tableau-Anfragen NP-vollständig ist (siehe Theorem 3.21 und Lemma 3.14). Es reicht also, eine Polynomialzeit-Reduktion f vom Auswertungsproblem für Boolesche Tableau-Anfragen (kurz: AWP) aufs Query Containment Problem für Tableau-Anfragen (kurz: QCP) zu finden.

Unsere Reduktion f bildet eine Instanz (\mathbf{I}, Q) des AWP auf die QCP-Instanz

$$f(\mathbf{I}, Q) := (P_{\mathbf{I}}, Q)$$

ab, wobei $P_{\mathbf{I}} := (\top_{\mathbf{I}}, v)$ mit $v := ()$ und $\top_{\mathbf{I}} := \mathbf{I}$.

Offensichtlicherweise sind $P_{\mathbf{I}}$ und Q Boolesche Tableau-Anfragen, die bei Eingabe von \mathbf{I} und Q in polynomieller Zeit konstruiert werden können.

Wir müssen also nur noch zeigen, dass Folgendes gilt:

$$\begin{array}{l} (\mathbf{I}, Q) \text{ ist eine „ja“-Instanz} \\ \text{fürs AWP, d.h. } () \in \llbracket Q \rrbracket(\mathbf{I}) \end{array} \iff \begin{array}{l} (P_{\mathbf{I}}, Q) \text{ ist eine „ja“-Instanz} \\ \text{fürs QCP, d.h. } P_{\mathbf{I}} \sqsubseteq Q. \end{array}$$

Gemäß Homomorphismussatz gilt:

$$\begin{array}{l} P_{\mathbf{I}} \sqsubseteq Q \iff \text{es gibt einen Homomorphismus von } Q \text{ auf } P_{\mathbf{I}} \\ \iff u_{P_{\mathbf{I}}}^Q \in \llbracket Q \rrbracket(\mathbf{I}_{P_{\mathbf{I}}}^Q). \end{array}$$

Hierbei ist $u_{P_{\mathbf{I}}}^Q = ()$. Da $\top_{\mathbf{I}}$ keine Variablen enthält und $\top_{\mathbf{I}} = \mathbf{I}$ ist, gilt außerdem: $\mathbf{I}_{P_{\mathbf{I}}}^Q = \top_{\mathbf{I}} = \mathbf{I}$. Also gilt:

$$P_{\mathbf{I}} \sqsubseteq Q \iff () \in \llbracket Q \rrbracket(\mathbf{I}).$$

Dies beendet den Beweis der NP-Härte des QCP. □

Bemerkung:

Für die im obigen Beweis konstruierte Tableau-Anfrage $P_{\mathbf{I}}$ gilt übrigens für jede Datenbank \mathbf{J} vom Schema \mathbf{S} : $() \in \llbracket P_{\mathbf{I}} \rrbracket(\mathbf{J}) \iff \mathbf{I} \subseteq \mathbf{J}$. D.h. bei Eingabe einer Datenbank \mathbf{J} liefert die Anfrage $P_{\mathbf{I}}$ genau dann die Antwort „ja“, wenn \mathbf{J} eine Erweiterung der Datenbank \mathbf{I} ist.

Bemerkung: Wegen

$$Q \equiv Q' \iff (Q \sqsubseteq Q' \text{ und } Q' \sqsubseteq Q)$$

gibt es laut Korollar 3.36 insbes. auch einen Algorithmus, der bei Eingabe zweier Tableau-Anfragen Q und Q' entscheidet, ob die beiden Anfragen äquivalent sind.

Tableau-Minimierung

Definition 3.37. Sei \mathbf{S} ein Datenbankschema.

- (a) Eine Tableau-Anfrage (\mathbb{T}, u) heißt *minimal*, falls es keine zu (\mathbb{T}, u) äquivalente Tableau-Anfrage (\mathbb{T}', u') mit $|\mathbb{T}'| < |\mathbb{T}|$ gibt (wobei $|\mathbb{T}|$ die Kardinalität, d.h. die Gesamtzahl von Tupeln in \mathbb{T} bezeichnet).
- (b) Zwei Tableau-Anfragen $Q' := (\mathbb{T}', u')$ und $Q := (\mathbb{T}, u)$ heißen *isomorph*, falls es eine Bijektion $\pi : \text{var}(Q') \rightarrow \text{var}(Q)$ gibt, so dass $\pi(\mathbb{T}') = \mathbb{T}$ und $\pi(u') = u$ ist.

Theorem 3.38 (Chandra, Merlin, 1977).

- (a) Zu jeder Tableau-Anfrage (\mathbb{T}, u) gibt es ein $\mathbb{T}' \subseteq \mathbb{T}$ (d.h. für jedes $R \in \mathbf{S}$ ist $\mathbb{T}'(R) \subseteq \mathbb{T}(R)$), so dass die Anfrage (\mathbb{T}', u) minimal und äquivalent zu (\mathbb{T}, u) ist.
- (b) Sind (\mathbb{T}_1, u_1) und (\mathbb{T}_2, u_2) zwei minimale äquivalente Tableau-Anfragen, so sind (\mathbb{T}_1, u_1) und (\mathbb{T}_2, u_2) isomorph.

Beweis.

(a): Sei $Q := (\mathbb{T}, u)$ die gegebene Tableau-Anfrage, und sei $Q_1 = (\mathbb{T}_1, u_1)$ eine *minimale* Tableau-Anfrage, die äquivalent zu Q ist. Dann ist $Q_1 \sqsubseteq Q$ und $Q \sqsubseteq Q_1$. Gemäß Homomorphismussatz gibt es also einen Homomorphismus h von Q auf Q_1 und einen Homomorphismus g von Q_1 auf Q .

Skizze:

$$\begin{array}{ccccc}
 Q & & Q_1 & & Q \\
 \parallel & & \parallel & & \parallel \\
 (\mathbb{T}, u) & \xrightarrow{h} & (\mathbb{T}_1, u_1) & \xrightarrow{g} & (\mathbb{T}, u)
 \end{array}$$

Betrachte die Abbildung $f := g \circ h$. Dann ist f eine Abbildung von $\text{Var}(Q)$ auf $\text{var} \cup \text{dom}$ mit

$$(1) f(u) = g(h(u)) \stackrel{h \text{ Hom.}}{=} g(u_1) \stackrel{g \text{ Hom.}}{=} u \quad \text{und}$$

$$(2) f(\mathbb{T}) = g(h(\mathbb{T})) \stackrel{h \text{ Hom.}}{\subseteq} g(\mathbb{T}_1) \stackrel{g \text{ Hom.}}{\subseteq} \mathbb{T}.$$

Sei $\mathbb{T}' := f(\mathbb{T})$ und sei $Q' := (\mathbb{T}', u)$. Wegen (2) ist $\mathbb{T}' \subseteq \mathbb{T}$. Um den Beweis von (a) abzuschließen, müssen wir noch Folgendes zeigen:

(i) $Q' \equiv Q$ und

(ii) Q' ist minimal.

Für (i) beachte, dass die Abbildung f ein Homomorphismus von Q auf Q' ist. Gemäß Homomorphismussatz ist also $Q' \sqsubseteq Q$. Umgekehrt ist die Abbildung id (mit $id(x) := x$ für alle $x \in \mathbf{var}$) ein Homomorphismus von Q' auf Q . Gemäß Homomorphismussatz ist also $Q \sqsubseteq Q'$. Insgesamt ist daher $Q \equiv Q'$.

Für (ii) beachte, dass gemäß (2) gilt: $\mathbf{T}' = f(\mathbf{T}) \subseteq g(\mathbf{T}_1)$. Insbesondere ist also $|\mathbf{T}'| \leq |\mathbf{T}_1|$. Da Q_1 eine *minimale* zu Q äquivalente Tableau-Anfrage ist und da Q' zu Q äquivalent ist, muss $|\mathbf{T}'| \geq |\mathbf{T}_1|$ sein. Insgesamt ist also $|\mathbf{T}'| = |\mathbf{T}_1|$, und somit ist Q' eine zu Q äquivalente *minimale* Tableau-Anfrage. Dies beendet den Beweis von (a).

(b): Ähnlich (Details: Übung). □

Folie 109

Theorem 3.39.

(a) Es gibt einen Algorithmus, der bei Eingabe einer Tableau-Anfrage $Q = (\mathbf{T}, u)$ eine *minimale* zu Q äquivalente Tableau-Anfrage (\mathbf{T}', u) mit $\mathbf{T}' \subseteq \mathbf{T}$ berechnet.

(b) Das Problem

Eingabe: Tableau-Anfrage (\mathbf{T}, u) und Tableau $\mathbf{T}' \subseteq \mathbf{T}$
 Frage: Ist $(\mathbf{T}, u) \equiv (\mathbf{T}', u)$?

ist NP-vollständig.

Beweis.

(a): Sei $Q = (\mathbf{T}, u)$ die eingegebene Tableau-Anfrage und sei \mathbf{S} das DB-Schema, über dem Q formuliert ist.

Für jedes $R \in \mathbf{S}$ sei $n_R := |\mathbf{T}(R)|$ die Anzahl der Tupel im Einzel-Tableau $\mathbf{T}(R)$, und sei $u_{R,1}, \dots, u_{R,n_R}$ eine Auflistung aller Tupel in $\mathbf{T}(R)$.

Sei $\ell := |\mathbf{S}|$, und sei R_1, \dots, R_ℓ eine Auflistung aller Relationsnamen in \mathbf{S} .

Unser Minimierungs-Algorithmus geht bei Eingabe von $Q = (\mathbf{T}, u)$ wie folgt vor:

- (1) Setze $\hat{T} := T$ und $\hat{Q} := (\hat{T}, u)$.
- (2) Für $j = 1, \dots, \ell$ tue Folgendes:
 - (3) Für $i = 1, \dots, n_{R_j}$ tue Folgendes:
 - (4)
 - Sei \hat{T}' das Tableau, das aus \hat{T} entsteht, indem das Tupel $u_{R_j, i}$ aus $\hat{T}(R_j)$ gelöscht wird.
 - Teste, ob für $\hat{Q}' := (\hat{T}', u)$ gilt: $\hat{Q}' \sqsubseteq \hat{Q}$.
 - Falls „ja“, so setze $\hat{T} := \hat{T}'$ und $\hat{Q} := \hat{Q}'$.
- (5) Gib $\hat{Q} := (\hat{T}, u)$ aus.

Wir müssen zeigen, dass für die vom Algorithmus ausgegebene Anfrage \hat{Q} gilt:

- (i) $\hat{Q} \equiv Q$ und
- (ii) \hat{Q} ist minimal.

An Stelle von (i) zeigen wir sogar, dass während des gesamten Laufs des Algorithmus \hat{Q} stets äquivalent zu Q ist. Zu Beginn des Algorithmus gilt dies, da $\hat{Q} = Q$ ist. Gemäß Induktionsannahme gelte nun zu Beginn des Schleifendurchlaufs für j und i , dass $\hat{Q} \equiv Q$ ist. Wir müssen zeigen, dass auch am Ende dieses Schleifendurchlaufs \hat{Q} äquivalent zu Q ist. Sei dazu \hat{Q}_0 die Anfrage \hat{Q} zu Beginn des Schleifendurchlaufs, und sei \hat{Q}_1 die Anfrage \hat{Q} am Ende des Schleifendurchlaufs.

Fall 1: Der Test während des Schleifendurchlaufs liefert die Antwort „nein“. Dann ist $\hat{Q}_1 = \hat{Q}_0$ und gemäß Induktionsannahme äquivalent zu Q .

Fall 2: Der Test während des Schleifendurchlaufs liefert die Antwort „ja“. Dann ist $\hat{Q}' \sqsubseteq \hat{Q}_0$. Außerdem ist das Tableau von \hat{Q}' im Tableau von \hat{Q}_0 enthalten. Daher ist die identische Abbildung id ein Homomorphismus von \hat{Q}' auf \hat{Q}_0 . Laut Homomorphismus-Satz ist also $\hat{Q}_0 \sqsubseteq \hat{Q}'$. Insgesamt erhalten wir: $\hat{Q}' \equiv \hat{Q}_0 \equiv Q$. Am Ende des Schleifendurchlaufs wird \hat{Q} durch \hat{Q}' ersetzt. Also ist am Ende des Schleifendurchlaufs $\hat{Q} \equiv Q$.

Für (ii) führen wir einen Beweis durch Widerspruch. Angenommen, \hat{Q} ist *nicht* minimal. Laut Theorem 3.38 gibt es ein Tableau $T_1 \subseteq \hat{T}$, so dass die Anfrage (T_1, u) minimal und äquivalent zu \hat{Q} ist.

Da laut Annahme \hat{Q} *nicht* minimal ist, ist $\hat{T} \neq T_1$.

Sei j_0 das kleinste $j \in \{1, \dots, \ell\}$, für das $\hat{T}(R_j) \neq T_1(R_j)$ ist.

Sei i_0 das kleinste $i \in \{1, \dots, n_{R_{j_0}}\}$, für das gilt:

$$u_{R_{j_0}, i} \in \hat{T}(R_{j_0}) \quad \text{und} \quad u_{R_{j_0}, i} \notin T_1(R_{j_0}). \quad (3.1)$$

Bei unserem Algorithmus ist zu Beginn des Schleifendurchlaufs für $j = j_0$ und $i = i_0$ dann \hat{T} das Tableau, bei dem für jedes $j' \in \{1, \dots, \ell\}$ gilt:

$$\hat{T}(R_{j'}) = \begin{cases} T_1(R_{j'}) & , \text{ falls } j' < j_0 \\ T_1(R_{j_0}) \cup \{u_{R_{j_0}, i'} : i' \geq i_0\} & , \text{ falls } j' = j_0 \\ T(R_{j'}) & , \text{ falls } j' > j_0. \end{cases}$$

Während des Schleifendurchlaufs für $j = j_0$ und $i = i_0$ ist dann \hat{T}' das Tableau, das aus \hat{T} entsteht, indem das Tupel $u_{R_{j_0}, i_0}$ aus $\hat{T}(R_{j_0})$ entfernt wird. Und es ist $\hat{Q}' = (\hat{T}', u)$.

Gemäß (3.1) ist $u_{R_{j_0}, i_0} \notin T_1(R_{j_0})$.

Somit ist $T_1 \subseteq \hat{T}'$. Daher ist die identische Abbildung id ein Homomorphismus von Q_1 auf \hat{Q}' . Gemäß Homomorphismus-Satz ist also $\hat{Q}' \sqsubseteq Q_1$. Und da $Q_1 \equiv Q$ ist, gilt: $\hat{Q}' \sqsubseteq Q$. Der im Algorithmus während des Schleifendurchlaufs für $j = j_0$ und $i = i_0$ durchgeführte Test auf $\hat{Q}' \sqsubseteq \hat{Q}$ liefert also die Antwort „ja“ (da, wie wir bereits gezeigt haben, $\hat{Q} \equiv Q$ gilt). Somit wird am Ende des Scheifendurchlaufs $\hat{T} := \hat{T}'$ gesetzt. Daher gilt für die am Ende vom Algorithmus ausgegebene Anfrage $\hat{Q} = (\hat{T}, u)$, dass $u_{R_{j_0}, i_0} \notin \hat{T}(R_{j_0})$ ist. Gemäß (3.1) müsste aber gelten: $u_{R_{j_0}, i_0} \in \hat{T}(R_{j_0})$. Widerspruch!

Dies beendet den Beweis von (ii) und damit auch den Beweis von (a).

(b): Übung. □

Optimierung von SPJR-Anfragen

Vorgehensweise bei Eingabe einer SPJR-Anfrage Q :

- (1) Übersetze Q in eine Tableau-Anfrage (T, u) (bzw. gib „unerfüllbar“ aus, falls Q nicht erfüllbar ist)
- (2) Finde minimales Tableau $T' \subseteq T$ so dass (T', u) äquivalent zu (T, u) ist.
- (3) Übersetze (T', u) in eine äquivalente SPJR-Anfrage Q'

- (4) Wende heuristische Optimierung (siehe Kapitel 6) auf Q' an und werte Q' aus.

Die Minimierung des Tableaus entspricht der Minimierung der Anzahl der Joins, denn: Anzahl Zeilen im Tableau = 1 + Anzahl Join-Operationen bei der Auswertung der Anfrage.

Beispiel 3.40.

- $S = \{R\}$, wobei R die Attribute A, B, C hat.
- $Q := \pi_{A,B}(\sigma_{B="5"}(R)) \bowtie \pi_{B,C}(\pi_{A,B}(R) \bowtie \pi_{A,C}(\sigma_{B="5"}(R)))$
- zugehörige Tableau-Anfrage: (T, u) mit $u = (x_A, 5, z_C)$ und

$$T(R) = \begin{array}{|c|c|c|} \hline x_A & 5 & x_C \\ y_A & 5 & y_C \\ y_A & 5 & z_C \\ \hline \end{array}$$

- minimales Tableau T' : Wir nutzen den Minimierungsalgorithmus aus dem Beweis von Theorem 3.39, um $Q = (T, u)$ zu minimieren.
 - Setze $\hat{T} := T$ und $\hat{Q} := (\hat{T}, u)$. Setze $j := 1$ und $R_1 := R$. Betrachte die Schleifendurchläufe für $i = 1, \dots, 3$.
 - Schleifendurchlauf für $i = 1$: Sei \hat{T}' das Tableau mit

$$\hat{T}'(R) = \begin{array}{|c|c|c|} \hline y_A & 5 & y_C \\ y_A & 5 & z_C \\ \hline \end{array}$$

Teste, ob für $\hat{Q}' := (\hat{T}', u)$ gilt: $\hat{Q}' \sqsubseteq \hat{Q}$:

Gemäß Homomorphismus-Satz genügt es dafür, zu testen, ob es einen Homomorphismus von \hat{Q} auf \hat{Q}' gibt. Angenommen, h wäre ein solcher Homomorphismus. Wegen $h(u) = u$ muss dann $h(x_A) = x_A$ und $h(z_C) = z_C$ sein. Die erste Zeile von $\hat{T}(R) = T(R)$ wird durch h dann abgebildet auf $(x_A, 5, h(x_C))$. Aber $\hat{T}'(R)$ enthält keine solche Zeile. Daher kann es keinen Homomorphismus von \hat{Q} auf \hat{Q}' geben. Es ist also $\hat{Q}' \not\sqsubseteq \hat{Q}$. Am Ende des Schleifendurchlaufs ist also immer noch $\hat{T} = T$.

- Schleifendurchlauf für $i = 2$: Sei \hat{T}' das Tableau mit

$$\hat{T}'(R) = \begin{array}{|c|c|c|} \hline x_A & 5 & x_C \\ \hline y_A & 5 & z_C \\ \hline \end{array}$$

Teste, ob für $\hat{Q}' := (\hat{T}', u)$ gilt: $\hat{Q}' \sqsubseteq \hat{Q}$:

Gemäß Homomorphismus-Satz genügt es dafür, zu testen, ob es einen Homomorphismus von \hat{Q} auf \hat{Q}' gibt. Angenommen, h wäre ein solcher Homomorphismus. Wegen $h(u) = u$ muss dann $h(x_A) = x_A$ und $h(z_C) = z_C$ sein. Wir wählen h so, dass außerdem gilt: $h(x_C) = x_C$ und $h(y_C) = z_C$. Man kann leicht nachprüfen, dass diese Abbildung h tatsächlich ein Homomorphismus von \hat{Q} auf \hat{Q}' ist. Es gilt also $\hat{Q}' \sqsubseteq \hat{Q}$. Am Ende des Schleifendurchlaufs wird also \hat{T} durch \hat{T}' ersetzt.

- Schleifendurchlauf für $i = 3$: Sei \hat{T}' das Tableau mit

$$\hat{T}'(R) = \begin{array}{|c|c|c|} \hline x_A & 5 & x_C \\ \hline \end{array}$$

Teste, ob für $\hat{Q}' := (\hat{T}', u)$ gilt: $\hat{Q}' \sqsubseteq \hat{Q}$:

Gemäß Homomorphismus-Satz genügt es dafür, zu testen, ob es einen Homomorphismus von \hat{Q} auf \hat{Q}' gibt. Angenommen, h wäre ein solcher Homomorphismus. Wegen $h(u) = u$ muss dann $h(x_A) = x_A$ und $h(z_C) = z_C$ sein. Zur Erinnerung: Jetzt ist

$$\hat{T}(R) = \begin{array}{|c|c|c|} \hline x_A & 5 & x_C \\ \hline y_A & 5 & z_C \\ \hline \end{array}$$

Die letzte Zeile von $\hat{T}(R)$ wird durch h dann abgebildet auf $(h(y_A), 5, z_C)$. Aber $\hat{T}'(R)$ enthält keine solche Zeile. Daher kann es keinen Homomorphismus von \hat{Q} auf \hat{Q}' geben. Es ist also $\hat{Q}' \not\sqsubseteq \hat{Q}$. Am Ende des Schleifendurchlaufs bleibt also \hat{T} unverändert.

- Am Ende gibt der Algorithmus die minimale Tableau-Anfrage (Γ', u) aus mit $u = (x_A, 5, z_C)$ und

$$\hat{T}'(R) = \begin{array}{|c|c|c|} \hline x_A & 5 & x_C \\ \hline y_A & 5 & z_C \\ \hline \end{array}$$

- zugehörige SPJR-Anfrage:

$$Q' := \pi_{A,B}(\sigma_{B="5"}(R)) \bowtie \pi_{B,C}(\sigma_{B="5"}(R)).$$

3.5 Azyklische Anfragen

Folie 111

Motivation

- *Ziel jetzt:* Teilklasse der Klasse der konjunktiven Anfragen, für die das Auswertungsproblem in Polynomialzeit lösbar ist (kombinierte Komplexität)
- \rightsquigarrow *azyklische konjunktive Anfragen*
- Wir werden in diesem Kapitel oft nur *Boolesche* Anfragen betrachten (... wenn wir die effizient auswerten können, dann können wir die Konstruktion aus dem Beweis von Theorem 3.20 benutzen, um auch Anfragen auszuwerten, deren Ergebnis die Stelligkeit ≥ 1 hat)
- *In der Literatur:* Verschiedene äquivalente Definitionen (bzw. Charakterisierungen) der azyklischen Booleschen konjunktiven Anfragen, etwa:
 - regelbasierte konjunktive Anfragen mit azyklischem Hypergraph
 - regelbasierte konjunktive Anfragen der Hyperbaum-Weite 1
 - *regelbasierte konjunktive Anfragen, die einen Join-Baum besitzen*
 - *Boolesche Semijoin-Anfragen*
 - *konjunktive Sätze des Guarded Fragment*

Folie 112

Beispiel

Beispiel-Datenbank mit Relationen

- T mit Attributen *Student, Kurs, Semester* T steht für „Teilnehmer“
- D mit Attributen *Prof, Kurs, Semester* D steht für „Dozent“
- E mit Attributen *Person1, Person2*
..... steht für „Person1 ist Elternteil von Person2“

Anfrage 1: Gibt es einen Dozenten, dessen Tochter/Sohn an irgendeinem Kurs teilnimmt?

Als regelbasierte konjunktive Anfrage $Q_1 :=$

$$Ans() \leftarrow E(x_P, x_S), D(x_P, x_K, x_Z), T(x_S, y_K, y_Z)$$

Auswertung als „Semijoin-Anfrage“:

$$\pi\left(\left(E(x_P, x_S) \bowtie D(x_P, x_K, x_Z)\right) \bowtie T(x_S, y_K, y_Z)\right)$$

Anfrage 2: Gibt es einen Studenten, der an einem Kurs teilnimmt, der von seinem/r Vater/Mutter veranstaltet wird?

Als regelbasierte konjunktive Anfrage $Q_2 :=$

$$Ans() \leftarrow E(x_P, x_S), D(x_P, x_K, x_Z), T(x_S, x_K, x_Z)$$

Auswertung: $\pi\left(E(x_P, x_S) \bowtie D(x_P, x_K, x_Z) \bowtie T(x_S, x_K, x_Z)\right)$

Auswertung durch eine „Semijoin-Anfrage“ ist nicht möglich.

Folie 113

Join-Bäume & Azyklische regelbasierte conj. Anfragen

Definition 3.41.

(a) Sei $Q := Ans(u) \leftarrow R_1(u_1), \dots, R_\ell(u_\ell)$ eine regelbasierte konjunktive Anfrage. Ein *Join-Baum* von Q ist ein *Baum* mit *Knotenmenge* $\{R_1(u_1), \dots, R_\ell(u_\ell)\}$, so dass für alle Knoten $R_i(u_i)$ und $R_j(u_j)$ die folgende „*Weg-Eigenschaft*“ gilt:

Jede Variable x , die sowohl in u_i als auch in u_j vorkommt, kommt in *jedem* Knoten vor, der auf dem (eindeutig bestimmten) Weg zwischen $R_i(u_i)$ und $R_j(u_j)$ liegt.

(b) Eine regelbasierte konjunktive Anfrage Q (beliebiger Stelligkeit) heißt *azyklisch*, falls es einen Join-Baum für Q gibt.

Beispiele:

- Join-Baum für $Q_1 := Ans() \leftarrow E(x_P, x_S), D(x_P, x_K, x_Z), T(x_S, y_K, y_Z)$
- Es gibt keinen Join-Baum für $Q_2 := Ans() \leftarrow E(x_P, x_S), D(x_P, x_K, x_Z), T(x_S, x_K, x_Z)$
- Join-Baum für $Q_3 := Ans() \leftarrow R(y, z), P(x, y), S(y, z, u), S(z, u, w), T(y, z), T(z, u), R(z', y')$

Folie 114

Effiziente Auswertung von azyklischen Booleschen konjunktiven Anfragen

Vorgehensweise:

Eingabe: Boolesche regelbasierte konjunktive Anfrage Q , Datenbank \mathbf{I}

Ziel: Berechne $\llbracket Q \rrbracket(\mathbf{I})$

(1) Teste, ob Q azyklisch ist und konstruiere ggf. einen Join-Baum T für Q .
(*Details dazu: später*)

(2) Nutze T zur Konstruktion einer Booleschen *Semijoin-Anfrage* Q' ,
die äquivalent zu Q ist.
(*Details dazu: gleich*)

(3) Werte Q' in \mathbf{I} aus
(*das geht gemäß Proposition 3.43 in Zeit $\mathcal{O}(\|Q'\| \cdot \|\mathbf{I}\| \cdot \log(\|\mathbf{I}\|))$*)

Folie 115

Semijoin-Anfragen

Definition 3.42. Sei \mathbf{S} ein Datenbankschema.

Die Klasse der *Semijoin-Anfragen über \mathbf{S}* ist induktiv wie folgt definiert:

(A) Jedes Relations-Atom $R(v_1, \dots, v_r)$, für $R \in \mathbf{S}$, $r := \text{ar}(R)$ und
 $v_1, \dots, v_r \in \mathbf{var} \cup \mathbf{dom}$ ist eine Semijoin-Anfrage der Sorte (v_1, \dots, v_r) .

Semantik: Für jede Datenbank $\mathbf{I} \in \text{inst}(\mathbf{S})$ ist $\llbracket R(v_1, \dots, v_r) \rrbracket(\mathbf{I}) :=$

$$\left\{ \begin{array}{l} \beta : (\{v_1, \dots, v_r\} \cap \mathbf{var}) \rightarrow \mathbf{dom} \text{ ist eine} \\ (\beta(v_1), \dots, \beta(v_r)) : \text{Belegung, so dass} \\ (\beta(v_1), \dots, \beta(v_r)) \in \mathbf{I}(R) \end{array} \right\}$$

(S) Sind Q_1 und Q_2 Semijoin-Anfragen der Sorten (v_1, \dots, v_r) und
 (v'_1, \dots, v'_s) , so ist $Q := (Q_1 \times Q_2)$ eine Semijoin-Anfrage der Sorte
 (v_1, \dots, v_r) .

Semantik: Für jede Datenbank $\mathbf{I} \in \text{inst}(\mathbf{S})$ ist $\llbracket Q \rrbracket(\mathbf{I}) :=$

$$\left\{ \begin{array}{l} \text{es gibt ein } (b_1, \dots, b_s) \in \llbracket Q_2 \rrbracket(\mathbf{I}), \\ \text{so dass} \\ (a_1, \dots, a_r) \in \llbracket Q_1 \rrbracket(\mathbf{I}) : \text{für alle } i, j \text{ mit } v_i = v'_j \in \mathbf{var} \\ \text{gilt: } a_i = b_j \end{array} \right\}$$

Eine *Boolesche Semijoin-Anfrage* über \mathbf{S} ist von der Form $\pi(Q)$, wobei Q eine Semijoin-Anfrage über \mathbf{S} ist.

Folie 116

Auswertung von Semijoin-Anfragen

Proposition 3.43. *Das Auswertungsproblem für Semijoin-Anfragen bzw. Boolesche Semijoin-Anfragen ist in Zeit $\mathcal{O}(k \cdot n \cdot \log n)$ lösbar, wobei k die Größe der eingegebenen Anfrage und n die Größe der eingegebenen Datenbank ist.*

Beweis: siehe Tafel

□

Folie 117

Semijoin-Anfragen vs. Join-Bäume

Lemma 3.44.

- (a) *Es gibt einen Algorithmus, der bei Eingabe einer Semijoin-Anfrage Q in Zeit $\mathcal{O}(\|Q\|)$ eine zu Q äquivalente azyklische regelbasierte konjunktive Anfrage Q' und einen Join-Baum für Q' berechnet.*
- (b) *Es gibt einen Algorithmus, der bei Eingabe einer azyklischen Booleschen regelbasierten konjunktiven Anfrage Q und eines Join-Baums T für Q in Zeit $\mathcal{O}(\|Q\|)$ eine zu Q äquivalente Boolesche Semijoin-Anfrage Q' berechnet.*

Beweis: (a): Übung. (b): siehe Tafel.

□

Folgerung: Mit azyklischen Booleschen regelbasierten konjunktiven Anfragen kann man genau dieselben Anfragefunktionen ausdrücken wie mit Booleschen Semijoin-Anfragen.

Vorsicht: Dies gilt nicht, wenn man an Stelle von *Booleschen Anfragen* Anfragen beliebiger Stelligkeit betrachtet.

Folie 118

Konstruktion eines Join-Baums

Lemma 3.45. *Es gibt einen Polynomialzeit-Algorithmus, der bei Eingabe einer regelbasierten konjunktiven Anfrage Q entscheidet, ob Q azyklisch ist und ggf. einen Join-Baum für Q konstruiert.*

Beweis:

Algorithmus: **Eingabe:** Anfrage Q der Form $Ans(u) \leftarrow R_1(u_1), \dots, R_\ell(u_\ell)$

(1) $V := \{R_1(u_1), \dots, R_\ell(u_\ell)\}$ Knotenmenge

(2) $E := \emptyset$ Kantenmenge

(3) alle Elemente von V sind *unmarkiert*

(4) alle Variablen sind *unmarkiert*

(5) *Wiederhole* so lange, bis sich nichts mehr ändert:

(5.1) Falls es *unmarkierte Knoten* $R_i(u_i)$ und $R_j(u_j)$ (mit $i \neq j$) gibt, so dass alle *unmarkierten Variablen* aus u_j in u_i vorkommen, so *markiere den Knoten* $R_j(u_j)$ und füge in E eine *Kante* zwischen $R_i(u_i)$ und $R_j(u_j)$ ein.

(5.2) *Markiere* sämtliche *Variablen* x , für die gilt:
 „Es gibt *genau einen unmarkierten Knoten*, in dem x vorkommt.“

(6) Falls es *nur noch einen unmarkierten Knoten* gibt, so gib (V, E) aus;
 sonst gib aus: „ Q ist nicht azyklisch“.

Einige Details zum Beweis von Lemma 3.45

Beispiel: Probelauf des Algorithmus für die Anfragen

- $Q_1 := Ans() \leftarrow E(x_P, x_S), D(x_P, x_K, x_Z), T(x_S, y_K, y_Z)$
- $Q_2 := Ans() \leftarrow E(x_P, x_S), D(x_P, x_K, x_Z), T(x_S, x_K, x_Z)$
- $Q_3 := Ans() \leftarrow R(y, z), P(x, y), S(y, z, u), S(z, u, w), T(y, z), T(z, u), R(z', y')$

Notation:

- *Zeitpunkt* t = Beginn des t -ten Durchlaufs durch Zeile (5)
- $w_1^t, \dots, w_{r_t}^t$: die zu Zeitpunkt t noch unmarkierten Knoten
- MV^t : Menge der zum Zeitpunkt t bereits markierten Variablen
- E^t : die Kantenmenge zum Zeitpunkt t

Folie 120

Die Korrektheit des Algorithmus folgt direkt aus den folgenden Behauptungen 1 & 2:

Behauptung 1: Zu jedem Zeitpunkt t gilt:

- (1) _{t} : E^t ist ein Wald aus Bäumen $T_1^t, \dots, T_{r_t}^t$, deren Wurzeln die Knoten $w_1^t, \dots, w_{r_t}^t$ sind.
- (2) _{t} : Jeder dieser Bäume erfüllt die *Weg-Eigenschaft*, d.h. für alle $i \in \{1, \dots, r_t\}$, alle Knoten $v, v' \in T_i^t$ und jede Variable x , die sowohl in v als auch in v' vorkommt, gilt: x kommt in jedem Knoten auf dem Weg zwischen v und v' vor.
- (3) _{t} : Jede *unmarkierte Variable* (d.h. jede Variable, die nicht zu MV^t gehört), die in einem Baum T_k^t vorkommt, kommt auch in dessen Wurzel w_k^t vor.
- (4) _{t} : Es gibt keine *markierte Variable* (d.h. aus MV^t), die in 2 verschiedenen Bäumen T_i^t und T_j^t vorkommt.

Beweis: Induktion nach t .

$t = 1$: klar. $t \mapsto t+1$: Nachrechnen (Übung).

Behauptung 2:

Wenn Q azyklisch ist, so endet der Algorithmus mit nur *einem* unmarkierten Knoten.

Beweis: Siehe Tafel. □

Folie 121

Auswertungskomplexität azyklischer konjunktiver Anfragen

Theorem 3.46 (Yannakakis, 1981). *Das*

AWP FÜR AZYKLISCHE REGELBASIERTE KONJ. ANFRAGEN

Eingabe: Regelbasierte konjunktive Anfrage Q und Datenbank \mathbf{I}

Aufgabe: Falls Q azyklisch ist, so berechne $\llbracket Q \rrbracket(\mathbf{I})$;
ansonsten gib „ Q ist nicht azyklisch“ aus.

kann in Zeit polynomiell in $\|Q\| + \|\mathbf{I}\| + \|\llbracket Q \rrbracket(\mathbf{I})\|$ gelöst werden.

Beweis:

- **Algorithmus für Boolesche Anfragen:** Nutze Lemma 3.45, Lemma 3.44 (b) und Proposition 3.43.
- **Algorithmus für Anfragen beliebiger Stelligkeit:** Nutze den Algorithmus für Boolesche Anfragen und die Konstruktion aus dem Beweis von Theorem 3.20. Beachte dabei, dass sämtliche Boolesche Anfragen, die zur Auswertung einer azyklischen Anfrage Q gestellt werden, denselben Join-Baum besitzen wie Q und daher insbesondere azyklisch sind. \square

Bemerkung: Es ist bekannt, dass das Auswertungsproblem für Boolesche azyklische regelbasierte konjunktive Anfragen vollständig ist für die Komplexitätsklasse LOGCFL (Gottlob, Leone, Scarcello, 1998).

Folie 122

Konjunktives Guarded Fragment GF(CQ)

Definition 3.47. Sei \mathbf{S} ein Datenbankschema.

Mit $\text{GF}(\text{CQ})[\mathbf{S}]$ bezeichnen wir die Menge aller Formeln des konjunktiven Kalküls $\text{CQ}[\mathbf{S}]$ (vgl. Definition 3.8), die zum Guarded Fragment $\text{GF}[\mathbf{S}]$ gehören, d.h. ... (Details: siehe Tafel)

Konjunktive Sätze des Guarded Fragment sind Formeln aus $\text{GF}(\text{CQ})[\mathbf{S}]$, die keine freie(n) Variable(n) besitzen.

Folie 123

Azyklische Boolesche Konjunktive Anfragen

Satz 3.48. *Die folgenden Anfragesprachen können genau dieselben Booleschen Anfragefunktionen ausdrücken:*

- (a) *azyklische Boolesche regelbasierte konjunktive Anfragen,*
- (b) *Boolesche Semijoin-Anfragen,*
- (c) *konjunktive Sätze des Guarded Fragment.*

Und jede Anfrage aus einer dieser Anfragesprachen kann in polynomieller Zeit in äquivalente Anfragen der anderen Sprachen übersetzt werden.

Gemäß Theorem 3.46 ist das Auswertungsproblem (kombinierte Komplexität) also für jede dieser Anfragesprachen in Polynomialzeit lösbar.

Beweis: (a) \iff (b): Lemma 3.44. (a) \iff (c): Übung. □

3.6 Mengen-Semantik vs. Multimengen-Semantik

Folie 124

Motivation

bisher: *Mengen-Semantik* (engl.: *set semantics*):

- DB-Relation = eine Menge von Tupeln
- Duplikate eines Tupels werden eliminiert

$$\{t\} = \{t, t\} = \{t, t, t\} = \dots$$

in SQL:

- keine Duplikat-Elimination bei Anfragen der Form
`SELECT * FROM ... WHERE ...`
- falls Duplikat-Elimination explizit gewünscht:
`SELECT DISTINCT * FROM ... WHERE ...`

betrachte jetzt: *Multimengen-Semantik* (engl.: *bag semantics*):

$$\{t\} \neq \{t, t\} \neq \{t, t, t\} \neq \dots$$

Folie 125

Beispiel

Datenbankschema:

- 2-stellige Relation *Hersteller* mit Attributen *Name* und *Ort*
- 2-stellige Relation *Bauteil* mit Attributen *Teil* und *Lager*

„Datenbank“ \mathbf{I}_F mit

$\mathbf{I}_F(\mathbf{Hersteller})$

<i>Name</i>	<i>Ort</i>
Boeing	Seattle
Boeing	New York
Airbus	Hamburg

Notation:

- $|(Boeing, Seattle) |_{\mathbf{I}_F(Hersteller)} = 1$
- $|(Boeing, New York) |_{\mathbf{I}_F(Hersteller)} = 1$
- $|(Airbus, Hamburg) |_{\mathbf{I}_F(Hersteller)} = 1$

$\mathbf{I}_F(\mathbf{Bauteil})$

<i>Teil</i>	<i>Lager</i>
Motor	Seattle
Motor	Seattle
Flügel	Portland
Cockpit	Seattle
Cockpit	Seattle
Cockpit	Seattle

Notation:

- $|(Motor, Seattle) |_{\mathbf{I}_F(Bauteil)} = 2$
- $|(Flügel, Portland) |_{\mathbf{I}_F(Bauteil)} = 1$
- $|(Cockpit, Seattle) |_{\mathbf{I}_F(Bauteil)} = 3$

Multimengen (Bags) und Multimengen-Datenbanken

Sei M eine Menge.

- Eine *Multimenge* B über M ist eine Abbildung $B : M \rightarrow \mathbb{N}_{\geq 0}$
- Notation: Für $a \in M$ schreibe $|a|_B$ an Stelle von $B(a)$
 $|a|_B = i$ bedeutet: das Element a kommt i -mal in der Multimenge B vor.
- B heißt *endlich*, falls die Menge $\{a \in M : |a|_B \neq 0\}$ endlich ist.
- Für Multimengen B und B' über M gilt:
 - $B =_b B' \iff |a|_B = |a|_{B'}$, für alle $a \in M$
 - $B \subseteq_b B' \iff |a|_B \leq |a|_{B'}$, für alle $a \in M$
 - Insbesondere gilt: $B =_b B' \iff (B \subseteq_b B' \text{ und } B' \subseteq_b B)$
 - $B \cup_b B' :=$
 die Multimenge B'' über M mit $|a|_{B''} := |a|_B + |a|_{B'}$, für alle $a \in M$

Definition 3.49. Sei \mathbf{S} ein Datenbankschema.

Eine *Multimengen-Datenbank* $\mathbf{I} \in \text{inst}_b(\mathbf{S})$ ordnet jedem Relationssymbol $R \in \mathbf{S}$ eine endliche Multimenge $\mathbf{I}(R)$ über $\text{dom}^{\text{ar}(R)}$ zu.

Folie 127

Anfragen mit Multimengen-Semantik

Beispiel:

SQL-Anfrage:

```
SELECT B1.Teil, B2.Teil
FROM Bauteil B1, Bauteil B2
WHERE B1.Lager = B2.Lager
```

regelbasiert:

$\text{Ans}(x, y) \leftarrow \text{Bauteil}(x, z), \text{Bauteil}(y, z)$

Auswertung der SQL-Anfrage in Datenbank mit

$\mathbf{I}_F(\mathbf{Bauteil})$

<i>Teil</i>	<i>Lager</i>
Motor	Seattle
Motor	Seattle
Flügel	Portland
Cockpit	Seattle
Cockpit	Seattle
Cockpit	Seattle

- bilde Kreuzprodukt $\mathbf{I}_F(\text{Bauteil}) \times \mathbf{I}_F(\text{Bauteil})$
(ohne Duplikatelimination)
- wähle die Tupel, in denen die Lager-Komponenten gleich sind
- streiche die Spalten mit den Lager-Komponenten

Liefert als Ergebnis die Multimenge M mit

- $|(Motor, Motor)|_M = 4 \quad |(Flügel, Flügel)|_M = 1$
 $|(Cockpit, Cockpit)|_M = 9$
- $|(Motor, Cockpit)|_M = 6 = |(Cockpit, Motor)|_M$

Folie 128

Konjunktive Anfragen mit Multimengen-Semantik

Multimengen-Semantik $\llbracket Q \rrbracket_b$:

Sei $Q := \text{Ans}(u) \leftarrow R_1(u_1), \dots, R_\ell(u_\ell)$ eine regelbasierte konjunktive Anfrage der Stelligkeit r über einem Datenbankschema \mathbf{S} .

- Eine *Belegung* β für Q ist, wie bisher, eine Abbildung
 $\beta : \text{var}(Q) \rightarrow \mathbf{dom}$.
- Die Auswertung von Q in einer Multimengen-Datenbank $\mathbf{I} \in \text{inst}_b(\mathbf{S})$ liefert als Ergebnis die Multimenge $\llbracket Q \rrbracket_b(\mathbf{I})$, so dass für alle Tupel $t \in \mathbf{dom}^r$ gilt:

$$|t|_{\llbracket Q \rrbracket_b(\mathbf{I})} := \sum_{\substack{\beta : \beta \text{ Belegung} \\ \text{für } Q \text{ mit } \beta(u)=t}} \left(|\beta(u_1)|_{\mathbf{I}(R_1)} \cdot |\beta(u_2)|_{\mathbf{I}(R_2)} \cdots |\beta(u_\ell)|_{\mathbf{I}(R_\ell)} \right)$$

Äquivalenz und Query Containment von Anfragen:

Seien Q und Q' zwei regelbasierte konjunktive Anfragen derselben Stelligkeit r über einem Datenbankschema \mathbf{S} .

- $Q \equiv_b Q' \quad : \iff \quad \llbracket Q \rrbracket_b(\mathbf{I}) =_b \llbracket Q' \rrbracket_b(\mathbf{I}), \text{ für alle } \mathbf{I} \in \text{inst}_b(\mathbf{S})$
- $Q \sqsubseteq_b Q' \quad : \iff \quad \llbracket Q \rrbracket_b(\mathbf{I}) \subseteq_b \llbracket Q' \rrbracket_b(\mathbf{I}), \text{ für alle } \mathbf{I} \in \text{inst}_b(\mathbf{S})$
- Klar: $Q \equiv_b Q' \iff (Q \sqsubseteq_b Q' \text{ und } Q' \sqsubseteq_b Q)$
 Insbes: Äquivalenz ist höchstens so schwer wie Query Containment.

Folie 129

Ergebnisse

Theorem 3.50 (Chaudhuri, Vardi, 1993 (hier ohne Beweis)).

(a) Seien Q und Q' regelbasierte konjunktive Anfragen derselben Stelligkeit über demselben Datenbankschema. Dann ist $Q \equiv_b Q'$ genau dann, wenn die zu Q und Q' gehörenden Tableau-Anfragen isomorph sind (im Sinne von Definition 3.37).

(b) Das Problem

ÄQUIVALENZ KONJ. ANFRAGEN BZGL. MULTIMENGEN-SEMANTIK
 Eingabe: regelbasierte konjunktive Anfragen Q und Q'
 Frage: Ist $Q \equiv_b Q'$?

ist genauso schwer wie das Graph-Isomorphie-Problem.

Folie 130

Folgerungen und eine offene Frage

- Das Äquivalenzproblem bzgl. Multimengen-Semantik liegt in NP und ist vermutlich nicht NP-vollständig (da vermutet wird, dass das Graph-Isomorphie-Problem nicht NP-hart ist).
 Somit ist Äquivalenz bzgl. Multimengen-Semantik vermutlich einfacher als Äquivalenz bzgl. Mengen-Semantik.
- Das Query Containment Problem bzgl. Multimengen-Semantik ist vermutlich schwerer als das Query Containment Problem bzgl. Mengen-Semantik.

Offene Forschungsfrage:

Bisher ist nicht bekannt, ob das Query Containment Problem für konjunktive Anfragen bzgl. Multimengen-Semantik überhaupt entscheidbar ist.

Folie 131

Konj. Anfragen mit \neq in Multimengen-Semantik

Konjunktive regelbasierte Anfragen mit \neq sind von der Form

$$Ans(u) \leftarrow R_1(u_1), \dots, R_\ell(u_\ell), U_1, \dots, U_m$$

wobei $R_1(u_1), \dots, R_\ell(u_\ell)$ Relations-Atome und U_1, \dots, U_m Ungleichungen der Form $x \neq y$ mit $x, y \in \mathbf{var} \cup \mathbf{dom}$ sind.

Theorem 3.51 (Jayram, Kolaitis, Vee, 2006 (hier ohne Beweis)).

Das Problem

QCP für konj. Anfragen mit \neq bzgl. Multimengen-Semantik
Eingabe: Q und Q' : regelbasierte konjunktive Anfragen mit \neq
Frage: Ist $Q \sqsubseteq_b Q'$?

ist nicht entscheidbar.

Folie 132

Ab jetzt wieder

— und bis zum Ende des Semesters —

Mengen-Semantik

Kapitel 4

Datalog

4.1 Syntax, Semantik und Auswertungskomplexität

Folie 133

Beispiel: Frankfurter U-Bahn-Netz

Hier vereinfacht: Eine Relation *U-Bahn-Netz* mit Attributen *Linie*, *Halt*, *nächsterHalt*

U-Bahn-Netz

<i>Linie</i>	<i>Halt</i>	<i>nächsterHalt</i>
U4	Bockenheimer Warte	Festhalle/Messe
U4	Festhalle/Messe	Hauptbahnhof
U4	Hauptbahnhof	Willy-Brandt-Platz
U4	Willy-Brandt-Platz	Dom/Römer
...
U7
U7	Kirchplatz	Leipziger Str.
U7	Leipziger Str.	Bockenheimer Warte
U7	Bockenheimer Warte	Westend
...

Anfrage:

Gib alle Stationen aus, die von “Bockenheimer Warte” aus ohne Umsteigen zu erreichen sind.

(1) mit max. 1 Zwischenhalt:

$$\left\{ (x_S) : \exists x_L \left(U\text{-Bahn-Netz}(x_L, \text{“Bockenheimer Warte”}, x_S) \vee \right. \right.$$

$$\left. \begin{array}{l} \exists x_Z (U\text{-Bahn-Netz}(x_L, \text{“Bockenheimer Warte”}, x_Z) \wedge \\ U\text{-Bahn-Netz}(x_L, x_Z, x_S)) \end{array} \right\}$$

(2) mit max. 2 Zwischenhalten: *analog*

(3) mit *beliebig vielen* Zwischenhalten ???

(In einem späteren Kapitel werden wir sehen: *Nicht ausdrückbar in Relationaler Algebra*)

Folie 134

Erreichbarkeits-Anfrage in SQL

Im SQL (ab SQL-99 Standard) kann die Anfrage “Gib alle Stationen aus, die von “Bockenheimer Warte” aus ohne Umsteigen zu erreichen sind” folgendermaßen ausgedrückt werden:

```
WITH RECURSIVE Erreichbar(Linie,Start,Ziel)
AS (
  SELECT Linie, Halt, NaechsterHalt
  FROM U-Bahn-Netz
  UNION ALL
  SELECT Erreichbar.Linie,
         Erreichbar.Start,
         U-Bahn-Netz.NaechsterHalt
  FROM Erreichbar, U-Bahn-Netz
  WHERE Erreichbar.Linie = U-Bahn-Netz.Linie AND
        Erreichbar.Ziel = U-Bahn-Netz.Halt
)
SELECT Ziel
FROM Erreichbar
WHERE Start='Bockenheimer Warte'
```

Folie 135

Erreichbarkeits-Anfrage in Datalog

Die Anfrage

“Gib alle Stationen aus, die von “Bockenheimer Warte” aus ohne Umsteigen zu erreichen sind”

kann in *Datalog* folgendermaßen ausgedrückt werden:

$$\begin{aligned} \text{Erreichbar}(L, S, Z) &\leftarrow \text{U-Bahn-Netz}(L, S, Z) \\ \text{Erreichbar}(L, S, Z) &\leftarrow \text{Erreichbar}(L, S, Z'), \text{U-Bahn-Netz}(L, Z', Z) \\ \text{Ans}(Z) &\leftarrow \text{Erreichbar}(L, \text{“Bockenheimer Warte”}, Z) \end{aligned}$$

Folie 136

Datalog: Syntax

Definition 4.1.

(a) Eine *Datalog-Regel* ist ein Ausdruck der Form

$$R_0(u_0) \leftarrow R_1(u_1), \dots, R_\ell(u_\ell)$$

wobei $\ell \geq 0$, $R_0, R_1, \dots, R_\ell \in \mathbf{rel}$ und u_0, u_1, \dots, u_ℓ freie Tupel der Stelligkeiten $\text{ar}(R_0), \text{ar}(R_1), \dots, \text{ar}(R_\ell)$ sind, so dass jede Variable, die in u_0 vorkommt, auch in mindestens einem der Tupel u_1, \dots, u_ℓ vorkommt.

(b) Ein *Datalog-Programm* P ist eine *endliche Menge von Datalog-Regeln*.

(c) Eine *Datalog-Anfrage* (P, R) besteht aus einem Datalog-Programm P und einem Relationsnamen R , der in P vorkommt.

Folie 137

Notation

- Wie üblich bezeichnen wir mit $\text{adom}(P)$ bzw. $\text{adom}(Q)$ die Menge der Konstanten, die in einem Datalog-Programm P bzw. einer Datalog-Anfrage Q vorkommen.

Für eine Datenbank \mathbf{I} ist $\text{adom}(P, \mathbf{I}) := \text{adom}(P) \cup \text{adom}(\mathbf{I})$ und $\text{adom}(Q, \mathbf{I}) := \text{adom}(Q) \cup \text{adom}(\mathbf{I})$.

- $R_0(u_0)$ heißt *Kopf* der Regel $R_0(u_0) \leftarrow R_1(u_1), \dots, R_\ell(u_\ell)$;
 $R_1(u_1), \dots, R_\ell(u_\ell)$ heißt *Rumpf* der Regel.
- $edb(P)$ bezeichnet die Menge der Relationsnamen, die ausschließlich im *Rumpf* von Regeln in P vorkommen (die sog. “extensionalen Prädikate von P ”).
- $idb(P)$ bezeichnet die Menge der Relationsnamen, die im *Kopf* mindestens einer Regel in P vorkommen (die sog. “intensionalen Prädikate von P ”).
- $sch(P) := edb(P) \dot{\cup} idb(P)$ heißt *Schema* von P .

Folie 138

Beispiel

$$P := \left\{ \begin{array}{l} Erreichbar(L, S, Z) \leftarrow U\text{-Bahn-Netz}(L, S, Z) , \\ Erreichbar(L, S, Z) \leftarrow Erreichbar(L, S, Z'), U\text{-Bahn-Netz}(L, Z', Z) , \\ Ans(Z) \leftarrow Erreichbar(L, \text{“Bockenheimer Warte”}, Z) \end{array} \right\}$$

ist ein Datalog-Programm mit

- $edb(P) = \{ U\text{-Bahn-Netz} \}$
- $idb(P) = \{ Erreichbar, Ans \}$
- $sch(P) = \{ U\text{-Bahn-Netz}, Erreichbar, Ans \}$

$Q_1 := (P, Ans)$ ist eine Datalog-Anfrage;

$Q_2 := (P, Erreichbar)$ ist noch eine Datalog-Anfrage;

$Q_3 := (P, U\text{-Bahn-Netz})$ ist noch eine Datalog-Anfrage.

Folie 139

Datalog: Semantik

Die Semantik von Datalog lässt sich auf verschiedene Weisen definieren:

- *Fixpunkt-Semantik:*
“Regeln schrittweise anwenden, bis sich nichts mehr ändert”
- *Modellbasierte Semantik:*
kleinste Datenbank über $sch(P)$, die alle “Regeln wahr macht”
- *Beweisbasierte Semantik:*
“Fakten, die sich herleiten lassen, sind im Ergebnis”

Glücklicherweise kann man zeigen, dass alle drei Ansätze zum gleichen Resultat führen.

Wir betrachten im Folgenden hauptsächlich die Fixpunkt-Semantik, die folgendermaßen definiert ist:

Folie 140

Der “immediate consequence”-Operator T_P

Definition 4.2. Sei P ein Datalog-Programm.

- (a) Für jedes $R \in idb(P)$ seien $Q_{R,1}, \dots, Q_{R,k_R}$ diejenigen Regeln aus P , in deren Kopf das Relationssymbol R steht, und seien $Q'_{R,1}, \dots, Q'_{R,k_R}$ die Regeln, die aus $Q_{R,1}, \dots, Q_{R,k_R}$ entstehen, indem im *Kopf* jeweils R durch das Relationssymbol Ans' ersetzt wird (mit $Ans' \notin sch(P)$ und $ar(Ans') = ar(R)$).
- (b) Der “immediate consequence”-Operator $T_P : inst(sch(P)) \rightarrow inst(sch(P))$ ist folgendermaßen definiert:
Für jedes $\mathbf{J} \in inst(sch(P))$ und jedes $R \in sch(P)$ ist

$$T_P(\mathbf{J})(R) := \begin{cases} \mathbf{J}(R) & \text{falls } R \in edb(P) \\ \llbracket Q'_{R,1} \rrbracket(\mathbf{J}) \cup \dots \cup \llbracket Q'_{R,k_r} \rrbracket(\mathbf{J}) & \text{falls } R \in idb(P) \end{cases}$$

Beispiel: Ist P das Datalog-Programm aus dem Beispiel der Erreichbarkeits-Anfrage, und besteht $\mathbf{J}(Erreichbar)$ aus allen Tupeln, die “mit max. i Zwischenhalten ohne Umsteigen zu erreichen sind”, so besteht $T_P(\mathbf{J})(Erreichbar)$ aus allen Tupeln, die “mit max. $i + 1$ Zwischenhalten ohne Umsteigen zu erreichen sind”.

Folie 141

Monotonie von T_P

Bemerkung: Eine alternative (und äquivalente) Definition von T_P :

$$T_P(\mathbf{J})(R) := \{ \text{unmittelbare } R\text{-Konsequenzen von } P \text{ über } \mathbf{J} \},$$

wobei ein Tupel t eine *unmittelbare R -Konsequenz von P über \mathbf{J}* ist, falls

- $R \in edb(P)$ und $t \in \mathbf{J}(R)$ oder
- $R \in idb(P)$ und es eine Regel der Form $R(u) \leftarrow R_1(u_1), \dots, R_\ell(u_\ell)$ in P und eine Belegung β gibt, so dass $\beta(u) = t$ und $\beta(u_i) \in \mathbf{J}(R_i)$, für alle $i \in \{1, \dots, \ell\}$.

Lemma 4.3. Für jedes Datalog-Programm P gilt:

Der Operator T_P ist monoton, d.h. für alle $\mathbf{J}, \mathbf{J}' \in inst(sch(P))$ mit $\mathbf{J} \subseteq \mathbf{J}'$ (d.h. $\mathbf{J}(R) \subseteq \mathbf{J}'(R)$ f.a. $R \in sch(P)$) gilt: $T_P(\mathbf{J}) \subseteq T_P(\mathbf{J}')$.

Beweis: leicht (Übung).

Folie 142

Der “Stufen”-Operator S_P

Definition 4.4. Sei P ein Datalog-Programm.

Der *Stufen-Operator* $S_P : inst(sch(P)) \rightarrow inst(sch(P))$ ist folgendermaßen definiert:

Für jedes $\mathbf{J} \in inst(sch(P))$ und jedes $R \in sch(P)$ ist

$$S_P(\mathbf{J})(R) := \mathbf{J}(R) \cup T_P(\mathbf{J})(R).$$

Folie 143

Fixpunkte

Bemerkung 4.5.

(a) Man sieht leicht, dass für jedes $\mathbf{J} \in inst(sch(P))$ gilt:

$$\mathbf{J} \subseteq S_P(\mathbf{J}) \subseteq S_P(S_P(\mathbf{J})) \subseteq \dots \subseteq S_P^i(\mathbf{J}) \subseteq S_P^{i+1}(\mathbf{J}) \subseteq \dots$$

wobei $S_P^0(\mathbf{J}) := \mathbf{J}$ und $S_P^{i+1}(\mathbf{J}) := S_P(S_P^i(\mathbf{J}))$.

- (b) Man sieht leicht, dass $\text{adom}(S_P(\mathbf{J})) \subseteq \text{adom}(P, \mathbf{J})$ und $\text{adom}(S_P^i(\mathbf{J})) \subseteq \text{adom}(P, \mathbf{J})$, f.a. $i \in \mathbb{N}$.
- (c) Da $\text{adom}(P, \mathbf{J})$ nur endlich viele Elemente besitzt, muss es in der Inklusionskette aus (a) ein $i_0 \in \mathbb{N}$ mit $S_P^{i_0}(\mathbf{J}) = S_P^{i_0+1}(\mathbf{J})$ geben. Offensichtlicherweise gilt dann: $S_P^{i_0}(\mathbf{J}) = S_P^j(\mathbf{J})$ f.a. $j \geq i_0$.

Folie 144

Notation

- Ein $\mathbf{J} \in \text{inst}(\text{sch}(P))$ heißt *Fixpunkt* von S_P , falls $S_P(\mathbf{J}) = \mathbf{J}$.
- $\text{Ab-Stufe}(P, \mathbf{J}) := \min\{i_0 : S_P^{i_0}(\mathbf{J}) = S_P^{i_0+1}(\mathbf{J})\}$ heißt “*Abschluss-Stufe*” von P auf \mathbf{J} .

Die einzelnen Instanzen $S_P^0(\mathbf{J}), S_P^1(\mathbf{J}), S_P^2(\mathbf{J})$ etc. werden “*Stufen des Fixpunktprozesses*” genannt.

$\text{Ab-Stufe}(P, \mathbf{J})$ gibt also diejenige Stufe an, ab der der Fixpunkt erreicht ist:

$$\mathbf{J} = S_P^0(\mathbf{J}) \subsetneq S_P^1(\mathbf{J}) \subsetneq \dots \subsetneq S_P^{\text{Ab-Stufe}(P, \mathbf{J})}(\mathbf{J}) = S_P^{\text{Ab-Stufe}(P, \mathbf{J})+1}(\mathbf{J}) = S_P^j(\mathbf{J})$$

f.a. $j \geq \text{Ab-Stufe}(P, \mathbf{J})$

- $S_P^\omega(\mathbf{J}) := S_P^{\text{Ab-Stufe}(P, \mathbf{J})}(\mathbf{J})$. Klar: $S_P^\omega(\mathbf{J})$ ist ein Fixpunkt von S_P .

Folie 145

Fixpunkt-Semantik von Datalog

Definition 4.6.

- (a) Sei P ein Datalog-Programm, $\mathbf{S} := \text{edb}(P)$. Jeder Datenbank $\mathbf{I} \in \text{inst}(\mathbf{S})$ ordnen wir die Datenbank $\hat{\mathbf{I}} \in \text{inst}(\text{sch}(P))$ mit

$$\hat{\mathbf{I}}(R) := \begin{cases} \mathbf{I}(R) & \text{falls } R \in \text{edb}(P) \\ \emptyset & \text{falls } R \in \text{idb}(P) \end{cases}$$

zu. Ausgewertet in \mathbf{I} liefert P die Datenbank

$$\llbracket P \rrbracket(\mathbf{I}) := S_P^\omega(\hat{\mathbf{I}}) \in \text{inst}(\text{sch}(P))$$

- (b) Eine Datalog-Anfrage $Q = (P, R)$ liefert auf einer Datenbank $\mathbf{I} \in \text{inst}(\text{edb}(P))$ die Relation

$$\llbracket Q \rrbracket(\mathbf{I}) := (\llbracket P \rrbracket(\mathbf{I}))(R).$$

Folie 146

Auswertungskomplexität

Bemerkung:

Ein einfacher Algorithmus, der bei Eingabe von P und \mathbf{I} das Resultat $\llbracket P \rrbracket(\mathbf{I})$ berechnet, kann folgendermaßen vorgehen:

- (1) $\mathbf{J} := \hat{\mathbf{I}}$
- (2) Berechne $\mathbf{J}' := S_P(\mathbf{J})$
- (3) Falls $\mathbf{J}' \neq \mathbf{J}$, so ($\mathbf{J} := \mathbf{J}'$, GOTO (2))
- (4) Gib \mathbf{J} aus

Datenkomplexität dieses Algorithmus: Polynomialzeit.

Folie 147

Fixpunkt-Semantik vs. Modellbasierte Semantik

Notation:

Für zwei Datenbanken $\mathbf{J}, \mathbf{J}' \in \text{inst}(\mathbf{S})$ bezeichnet $\mathbf{J} \cap \mathbf{J}'$ die Datenbank mit $(\mathbf{J} \cap \mathbf{J}')(R) := \mathbf{J}(R) \cap \mathbf{J}'(R)$, f.a. $R \in \mathbf{S}$.

Für eine Menge $M \subseteq \text{inst}(\mathbf{S})$ ist $\bigcap M := \bigcap_{\mathbf{J} \in M} \mathbf{J}$.

Satz 4.7 (Knaster und Tarski).

Sei P ein Datalog-Programm und sei $\mathbf{J} \in \text{inst}(\text{sch}(P))$. Dann gilt:

$$S_P^\omega(\mathbf{J}) = \bigcap \{ \mathbf{J}' \in \text{inst}(\text{sch}(P)) : S_P(\mathbf{J}') = \mathbf{J}' \text{ und } \mathbf{J} \subseteq \mathbf{J}' \}.$$

D.h.: $S_P^\omega(\mathbf{J})$ ist die kleinste Erweiterung von \mathbf{J} , die ein Fixpunkt von S_P ist.

Beweis: Siehe Tafel.

Folie 148

Beweisbasierte Semantik von Datalog

Sichtweise:

- Ein *Faktum* ist ein Ausdruck der Form $R(t)$ mit $R \in \mathbf{rel}$ und $t \in \mathbf{dom}^{\text{ar}(R)}$.
- Eine Datenbank $\mathbf{I} \in \text{inst}(\mathbf{S})$ wird mit der folgenden *Menge von Fakten* identifiziert:

$$\text{Fakten}(\mathbf{I}) := \{ R(t) \ : \ R \in \mathbf{S} \text{ und } t \in \mathbf{I}(R) \}$$

- Für die beweisbasierte Semantik werden Datalog-Regeln als Schlussregel-Muster in Beweisen betrachtet.

Folie 149

Beweisbäume

Definition 4.8. Sei P ein Datalog-Programm und $\mathbf{J} \in \text{inst}(\text{sch}(P))$. Ein *Beweisbaum für ein Faktum $R(t)$ bzgl. \mathbf{J} und P* ist ein gerichteter Baum mit den folgenden Eigenschaften:

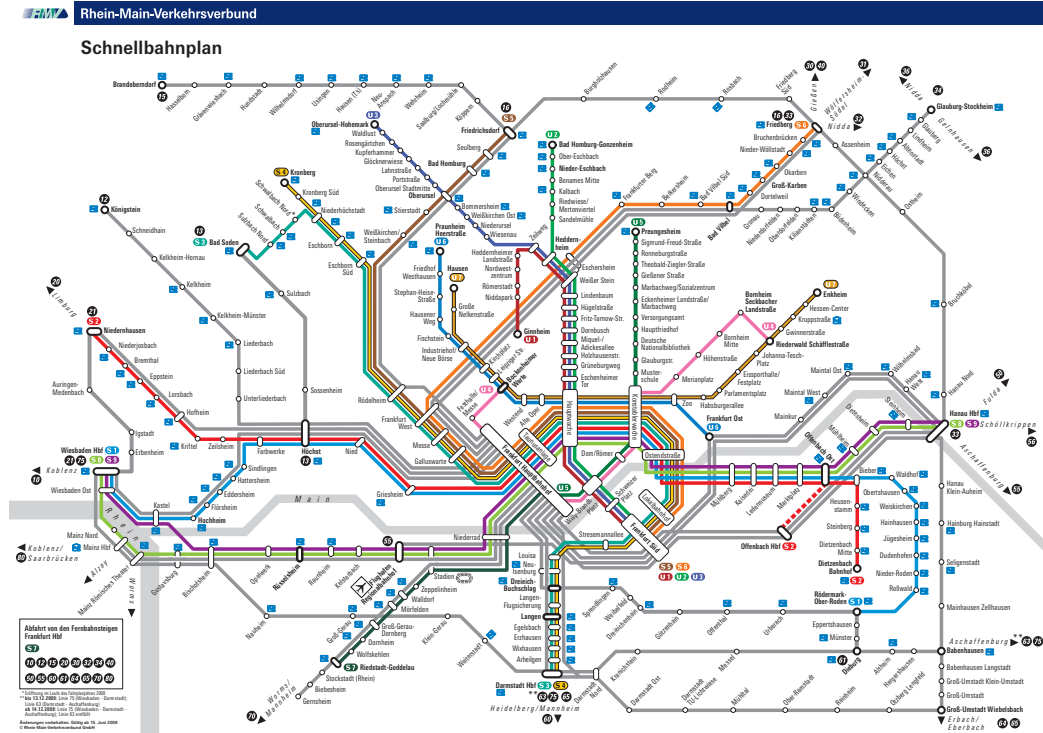
- (1) Jeder Knoten ist mit einem Faktum markiert.
- (2) Die Wurzel enthält das Faktum $R(t)$.
- (3) Die Fakten an den Blättern sind Elemente aus $\text{Fakten}(\mathbf{J})$
- (4) Für jeden inneren Knoten v mit Kindern v_1, \dots, v_ℓ gibt es eine Regel $R_0(u_0) \leftarrow R_1(u_1), \dots, R_\ell(u_\ell)$ in P und eine Belegung $\beta : \text{var}(P) \rightarrow \mathbf{dom}$ so dass gilt:
 - der Knoten v enthält das Faktum $R_0(\beta(u_0))$
 - das Kind v_i von v enthält das Faktum $R_i(\beta(u_i))$ (für alle $i \in \{1, \dots, \ell\}$).

Beispiel: Beweisbäume für $\text{Ans}(\text{“Frankfurt Süd”})$ bzgl. dem Frankfurter U-Bahn-Netz und dem folgenden Datalog-Programm:

$$\begin{aligned} E(L, S, Z) &\leftarrow U\text{-Bahn-Netz}(L, S, Z) \\ E(L, S, Z) &\leftarrow E(L, S, Y), U\text{-Bahn-Netz}(L, Y, Z) \\ \text{Ans}(Z) &\leftarrow E(L, \text{“Hauptwache”}, Z) \end{aligned}$$

(siehe Tafel)

Folie 150



Folie 151

Fixpunkt-Semantik vs. Beweisbasierte Semantik

Satz 4.9. Für jedes Datalog-Programm P , alle $\mathbf{J} \in \text{inst}(\text{sch}(P))$ und alle Fakten $R(t)$ mit $R \in \text{sch}(P)$ und $t \in \text{dom}^{\text{ar}(R)}$ gilt:

$$t \in S_P^\omega(\mathbf{J})(R) \iff \text{es gibt einen Beweisbaum für } R(t) \text{ bzgl. } \mathbf{J} \text{ und } P.$$

Beweis: Übung.

4.2 Grenzen der Ausdrucksstärke von Datalog

Folie 152

Monotonie und Abschluss unter $\text{adom}(Q)$ -Homomorphismen

Satz 4.10. Für jede Datalog-Anfrage Q gilt: $\llbracket Q \rrbracket$ ist monoton.

Beweis: Übung (folgt leicht aus der Monotonie des T_P -Operators).

Auf ähnliche Art lässt sich zeigen:

Satz 4.11. *Für jede Datalog-Anfrage Q gilt:
[[Q]] ist abgeschlossen unter $\text{adom}(Q)$ -Homomorphismen.*

Zur Erinnerung: Auf Übungsblatt 1 wurde Folgendes definiert.

- Ein C -Homomorphismus (für $C \subseteq \mathbf{dom}$) ist eine Abbildung $h : \mathbf{dom} \rightarrow \mathbf{dom}$ mit $h|_C = \text{id}$.
- Eine Anfragefunktion q ist *abgeschlossen unter C -Homomorphismen*, falls für alle C -Homomorphismen h und alle Datenbanken \mathbf{I} und \mathbf{J} gilt:

$$\text{Falls } h(\mathbf{I}) \subseteq \mathbf{J}, \text{ so ist } h(q(\mathbf{I})) \subseteq q(\mathbf{J}).$$

Folie 153

Ausdrucksstärke von Datalog

Bemerkung 4.12. Die Ausdrucksstärken von Datalog-Anfragen und von Anfragen der Relationalen Algebra sind unvergleichbar:

- Beispiel für eine Datalog-Anfrage, die nicht in Relationaler Algebra formuliert werden kann:
“Welche Stationen sind von “Bockenheimer Warte” aus ohne Umsteigen zu erreichen?”

- Beispiel für eine Relationale Algebra-Anfrage, die nicht in Datalog formuliert werden kann:
“In welchen Kinos läuft kein Film um 17:00 Uhr?”

Bzw. jede andere Anfrage, die nicht monoton ist (aber in Relationaler Algebra formuliert werden kann).

Aus der Monotonie von Datalog-Anfragen folgt leicht, dass diese Anfrage nicht in Datalog ausgedrückt werden kann.

4.3 Datalog zur Simulation von Turingmaschinen

Folie 154

Repräsentation von Worten als Datenbanken

Wir repräsentieren Worte über einem Alphabet Σ durch Datenbanken über dem Schema

$$\mathbf{S}_\Sigma := \{\text{Succ}, \text{Min}, \text{Max}\} \cup \{P_a : a \in \Sigma\},$$

wobei Succ 2-stellig ist und alle anderen Relationsnamen 1-stellig sind.

Für ein Wort $w = w_0 \cdots w_{n-1} \in \Sigma^*$ mit $|w| = n \geq 1$ und $w_0, \dots, w_{n-1} \in \Sigma$ ist für jedes $a \in \Sigma$

$$\mathbf{I}_w(P_a) := \{p \in \{0, \dots, n-1\} : w_p = a\}$$

und

$$\begin{aligned} \mathbf{I}_w(\text{Min}) &:= \{0\} \\ \mathbf{I}_w(\text{Max}) &:= \{n-1\} \\ \mathbf{I}_w(\text{Succ}) &:= \{(p, p+1) : 0 \leq p < n-1\}. \end{aligned}$$

Folie 155

Datalog zur Simulation von Turingmaschinen

Satz 4.13. *Für jede deterministische Turingmaschine M mit Eingabealphabet Σ und jede Zahl $k \geq 1$ gibt es ein Datalog-Programm $P_{M,k}$ mit $\text{edb}(P_{M,k}) = \mathbf{S}_\Sigma$ und ein 0-stelliges idb-Prädikat Ans von $P_{M,k}$, so dass für die Anfrage $Q_{M,k} := (P_{M,k}, \text{Ans})$ und für jedes Wort $w \in \Sigma^+$ gilt:*

$$\llbracket Q_{M,k} \rrbracket(\mathbf{I}_w) = \text{“yes”} \iff \text{bei Eingabe } w \text{ hält } M \text{ nach höchstens } |w|^k - 1 \text{ Schritten in einem akzeptierenden Zustand an.}$$

Beweis: Siehe Tafel.

Folie 156

Die Auswertungskomplexität von Datalog-Anfragen

Theorem 4.14 (Immerman und Vardi).

- (a) Die kombinierte Komplexität des Auswertungsproblems für Datalog-Anfragen ist EXPTIME-vollständig.
- (b) Die Datenkomplexität des Auswertungsproblems für Datalog-Anfragen ist PTIME-vollständig.

Beweis: Übung (unter Verwendung von ähnlichen Methoden wie beim Beweis von Satz 4.13).

Mit Aussage (b) ist Folgendes gemeint:

- (1) Für jede Datalog-Anfrage $Q = (P, R)$ ist das Problem

AWP_Q

Eingabe: Datenbank $\mathbf{I} \in \text{inst}(\text{edb}(P))$

Aufgabe: Berechne $\llbracket Q \rrbracket(\mathbf{I})$

in Zeit polynomiell in der Größe von \mathbf{I} lösbar, und

- (2) es gibt eine Boolesche Datalog-Anfrage Q , für die das Problem AWP_Q PTIME-vollständig ist (bzgl. logspace-Reduktionen).

4.4 Statische Analyse

Folie 157

Erfüllbarkeit

Theorem 4.15. *Das*

ERFÜLLBARKEITSPROBLEM FÜR DATALOG-ANFRAGEN

Eingabe: Datalog-Anfrage $Q = (P, R)$

Frage: Gibt es ein $\mathbf{I} \in \text{inst}(\text{edb}(P))$ so dass $\llbracket Q \rrbracket(\mathbf{I}) \neq \emptyset$?

ist entscheidbar.

Beweisidee:

Verallgemeinerung des Beweises der Erfüllbarkeit regelbasierter konjunktiver Anfragen (Satz 3.6).

Details: Übung.

Folie 158

Query Containment — Zwei Varianten

Herkömmliches *Query Containment*:

QUERY CONTAINMENT PROBLEM FÜR DATALOG-ANFRAGEN

Eingabe: Zwei Datalog-Anfragen $Q_1 = (P_1, R)$ und $Q_2 = (P_2, R)$ mit
 $edb(P_1) = edb(P_2)$ und $R \in idb(P_1) \cap idb(P_2)$

Frage: Gilt $\llbracket Q_1 \rrbracket(\mathbf{I}) \subseteq \llbracket Q_2 \rrbracket(\mathbf{I})$ für alle $\mathbf{I} \in inst(edb(P_1))$?

Uniformes Containment:

UNIFORMES CONTAINMENT PROBLEM FÜR DATALOG-PROGRAMME

Eingabe: Zwei Datalog-Programme P_1 und P_2 mit
 $edb(P_1) = edb(P_2)$ und $idb(P_1) = idb(P_2)$

Frage: Gilt $S_{P_1}^\omega(\mathbf{J}) \subseteq S_{P_2}^\omega(\mathbf{J})$ für alle $\mathbf{J} \in inst(sch(P_1))$?

Theorem 4.16.

- (a) Das *Query Containment Problem für Datalog-Anfragen* ist unentscheidbar.
- (b) Das *uniforme Containment Problem für Datalog-Programme* ist entscheidbar.

Beweis: Hier nur der Beweis von (a): siehe Tafel.

Folie 159

Beschränktheit (Boundedness)

- Wir haben gesehen: $\llbracket P \rrbracket(\mathbf{I})$ kann berechnet werden, indem man nach und nach

$$\hat{\mathbf{I}}, S_P(\hat{\mathbf{I}}), S_P(S_P(\hat{\mathbf{I}})), \dots, S_P^{Ab\text{-Stufe}(P, \mathbf{I})}(\hat{\mathbf{I}}), S_P^{Ab\text{-Stufe}(P, \mathbf{I})+1}(\hat{\mathbf{I}})$$

berechnet.

- Anzahl der Iterationen, die dafür nötig sind: $Ab\text{-Stufe}(P, \mathbf{I}) + 1$

- *Frage:* Wie groß kann $Ab\text{-Stufe}(P, \mathbf{I})$ werden?
- *Klar:* – in jeder Iteration kommt mindestens ein Faktum dazu
– $\text{adom}(\llbracket P \rrbracket(\mathbf{I})) \subseteq \text{adom}(P, \mathbf{I})$
- *Daher:* $Ab\text{-Stufe}(P, \mathbf{I}) \leq \sum_{R \in \text{idb}(P)} |\text{adom}(P, \mathbf{I})|^{\text{ar}(R)}$
- Besonders “schön” sind solche Datalog-Programme P , bei denen $Ab\text{-Stufe}(P, \mathbf{I})$ gar nicht von \mathbf{I} abhängt. Solche Programme heißen *beschränkt* (engl.: *bounded*).
- *Präzise:* Ein Datalog-Programm P heißt *beschränkt*, falls eine Zahl $d \in \mathbb{N}$ gibt, so dass für alle $\mathbf{I} \in \text{inst}(\text{edb}(P))$ gilt: $Ab\text{-Stufe}(P, \mathbf{I}) \leq d$. Den kleinsten solchen Wert d nennen wir *maximale Rekursionstiefe* von P .
- Wenn man weiß, dass ein Programm P beschränkt ist und maximale Rekursionstiefe d hat, so kann man leicht eine zu P äquivalente Anfrage in relationaler Algebra konstruieren, die nur die Operatoren der SPC-Algebra und den Vereinigungs-Operator benutzt (kurz: SPCU-Algebra).

Folie 160

Beispiel

Das Datalog-Programm

$$\begin{aligned} E(L, S, Z) &\leftarrow U\text{-Bahn-Netz}(L, S, Z) \\ E(L, S, Z) &\leftarrow E(L, S, Y), U\text{-Bahn-Netz}(L, Y, Z) \\ \text{Ans}(Z) &\leftarrow E(L, \text{“Bockenheimer Warte”}, Z) \end{aligned}$$

ist nicht beschränkt.

Das Datalog-Programm

$$\begin{aligned} \text{Kauft}(x, y) &\leftarrow \text{Mag}(x, y) \\ \text{Kauft}(x, y) &\leftarrow \text{Person}(x), \text{Trendsetter}(z), \text{Kauft}(z, y) \end{aligned}$$

ist beschränkt mit maximaler Rekursionstiefe 3, da für jede DB \mathbf{I} gilt:
 $T_P^\omega(\hat{\mathbf{I}})(\text{Kauft}) = \mathbf{I}(\text{Mag}) \cup \{(x, y) : x \in \mathbf{I}(\text{Person}), \text{ ex. } z \in$

$\mathbf{I}(\text{Trendsetter})$ s.d. $(z, y) \in \mathbf{I}(\text{Mag})\}$ und äquivalent zum nicht-rekursiven Programm

$$\begin{aligned} \text{Kauft}'(x, y) &\leftarrow \text{Mag}(x, y) \\ \text{Kauft}'(x, y) &\leftarrow \text{Person}(x), \text{Trendsetter}(z), \text{Mag}(z, y) \\ \text{Kauft}(x, y) &\leftarrow \text{Kauft}'(x, y) \\ \text{Kauft}(x, y) &\leftarrow \text{Person}(x), \text{Trendsetter}(z), \text{Kauft}'(z, y) \end{aligned}$$

Folie 161

Beschränktheit (Boundedness)

Theorem 4.17. *Das*

BOUNDEDNESS PROBLEM FÜR DATALOG-PROGRAMME

Eingabe: Datalog-Programm P

Frage: Ist P beschränkt, d.h. gibt es ein $d \in \mathbb{N}$ so dass
 $Ab\text{-Stufe}(P, \mathbf{I}) \leq d$ für alle $\mathbf{I} \in \text{inst}(\text{edb}(P))$?

ist unentscheidbar.

Beweis: siehe Tafel.

4.5 Einschränkung und Erweiterungen: nr-Datalog und Datalog mit Negation

Folie 162

Nicht-Rekursives Datalog — Beispiel

Beispiel 4.18. (a) Das Datalog-Programm

$$\begin{aligned} \text{Kauft}(x, y) &\leftarrow \text{Mag}(x, y) \\ \text{Kauft}(x, y) &\leftarrow \text{Person}(x), \text{Trendsetter}(z), \text{Mag}(z, y) \end{aligned}$$

ist nicht-rekursiv.

(b) Äquivalent dazu, aber NICHT nicht-rekursiv ist

$$\begin{aligned} \text{Kauft}(x, y) &\leftarrow \text{Mag}(x, y) \\ \text{Kauft}(x, y) &\leftarrow \text{Person}(x), \text{Trendsetter}(z), \text{Kauft}(z, y) \end{aligned}$$

Folie 163

Nicht-Rekursives Datalog — Präzise

Definition 4.19. Sei P ein Datalog-Programm.

- (a) Der *Abhängigkeitsgraph* G_P von P ist der gerichtete Graph mit Knotenmenge $V_P := sch(P)$ und Kantenmenge

$$E_P := \left\{ (R, S) : \begin{array}{l} \text{es gibt eine Regel in } P, \text{ in deren Rumpf } \\ R \text{ und in deren Kopf } S \text{ vorkommt} \end{array} \right\}$$

- (b) Die Klasse *nr-Datalog* aller *nicht-rekursiven Datalog-Programme* besteht aus allen Datalog-Programmen P , deren *Abhängigkeitsgraph* G_P *azyklisch* ist, d.h. keinen einfachen Kreis enthält.

Beispiele für Abhängigkeitsgraphen: siehe Tafel

Folie 164

“Nicht-rekursiv” vs. “beschränkt” (“bounded”)

Proposition 4.20. (a) *Jedes nr-Datalog-Programm ist beschränkt.*

- (b) *Jedes beschränkte Datalog-Programm ist äquivalent zu einem nr-Datalog-Programm.*

Beweis: Einfache Übung.

Folie 165

Ausdrucksstärke von nr-Datalog

- Die *SPCU-Algebra* ist die Erweiterung der SPC-Algebra um den *Vereinigungsoperator* \cup , der es erlaubt, die Ergebnisse zweier Anfragen derselben Stelligkeit zu vereinigen. Semantik:
 $\llbracket (Q_1 \cup Q_2) \rrbracket(\mathbf{I}) := \llbracket Q_1 \rrbracket(\mathbf{I}) \cup \llbracket Q_2 \rrbracket(\mathbf{I})$.
- Der *positive existentielle Kalkül* $PE-CALC_{\text{adom}}$ ist die Klasse aller Anfragen Q der Form $\{(e_1, \dots, e_r) : \varphi\}$, wobei φ eine Formel der Logik erster Stufe ist, in der keins der Symbole $\neg, \forall, \rightarrow, \leftrightarrow$ vorkommt. Semantik: $\llbracket Q \rrbracket(\mathbf{I}) := \{ \beta((e_1, \dots, e_r)) : \beta : \text{var} \rightarrow \text{adom}(Q, \mathbf{I}), \text{ so dass } \mathbf{I} \models \varphi[\beta] \}$

Satz 4.21. *Die folgenden Anfragesprachen können genau dieselben Anfragefunktionen beschreiben:*

(a) *nr-Datalog-Anfragen*

(b) *SPCU-Algebra*

(c) *positiver existentieller Kalkül $PE-CALC_{adom}$*

Bemerkung: Die Übersetzung von nr-Datalog in eine der anderen Sprachen kann mehr als polynomiell viel Zeit beanspruchen, da nr-Datalog-Programme u.U. viel kürzer sind als äquivalente Anfragen der anderen Sprachen.

Folie 166

Beweis von Satz 4.21

(b) \Rightarrow (a): Induktion über den Aufbau von SPCU-Anfragen (leicht).
Details: siehe Tafel.

(a) \Rightarrow (c): siehe Tafel.

(c) \Rightarrow (b): Bei gegebener Anfrage $Q = \{(e_1, \dots, e_r) : \varphi\}$ bringe die Formel φ zunächst in disjunktive Normalform, d.h. in eine Formel der Form $\varphi_1 \vee \dots \vee \varphi_\ell$, wobei φ_i eine $CQ^=$ -Formel ist.

Dann gilt für jede Datenbank \mathbf{I} , dass $\llbracket Q \rrbracket(\mathbf{I}) = \llbracket Q_1 \rrbracket(\mathbf{I}) \cup \dots \cup \llbracket Q_\ell \rrbracket(\mathbf{I})$, wobei

$$\llbracket Q_i \rrbracket(\mathbf{I}) = \{ \beta((e_1, \dots, e_r)) : \beta : \text{var} \rightarrow \text{adom}(Q, \mathbf{I}), \text{ so dass } \mathbf{I} \models \varphi_i \}.$$

Wir können nun ähnlich wie in Theorem 3.30 beim Beweis der Äquivalenz von Konjunktivem Kalkül und SPC-Algebra vorgehen, um eine SPCU-Anfrage Q'_i zu konstruieren, so dass für alle Datenbanken \mathbf{I} gilt:
 $\llbracket Q'_i \rrbracket(\mathbf{I}) = \llbracket Q_i \rrbracket(\mathbf{I})$.

Details: Übung. □

Folie 167

Datalog mit Negation

Ziel: Auch Negationszeichen “ \neg ” in Datalog-Regeln zulassen.

Definition 4.22. (a) Ein *Literal* ist ein Relationsatom $R(u)$ oder ein negiertes Relationsatom $\neg R(u)$.

Ein Literal der Form $R(u)$ heißt *positiv*; ein Literal der Form $\neg R(u)$ heißt *negativ* bzw. *negiert*.

(b) Eine *Datalog⁻-Regel* ist ein Ausdruck der Form

$$R_0(u_0) \leftarrow L_1(u_1), \dots, L_\ell(u_\ell)$$

wobei $\ell \geq 0$, $R_0 \in \mathbf{rel}$, u_0 ein freies Tupel der Stelligkeit $\text{ar}(R_0)$ und $L_1(u_1), \dots, L_\ell(u_\ell)$ Literale, so dass jede Variable, die in u_0 vorkommt, auch in mindestens einem *positiven* Literal $L_i(u_i)$ vorkommt.

(c) Ein *Datalog⁻-Programm* P ist eine *endliche Menge* von *Datalog⁻-Regeln*.

(d) Eine *Datalog⁻-Anfrage* (P, R) besteht aus einem *Datalog⁻-Programm* P und einem Relationsnamen R , der in P vorkommt.

Folie 168

Frage: Was soll die Semantik von *Datalog⁻* sein?

Beispiel 4.23. *Anfrage:*

“Gib alle Stationen aus, die von “Bockenheimer Warte” aus nicht ohne Umsteigen zu erreichen sind.”

Als Datalog⁻-Anfrage:

$$\begin{aligned} E(L, S, Z) &\leftarrow U\text{-Bahn-Netz}(L, S, Z) \\ E(L, S, Z) &\leftarrow E(L, S, Y), U\text{-Bahn-Netz}(L, Y, Z) \\ \text{Erreichbar_BW}(Z) &\leftarrow E(L, \text{“Bockenheimer Warte”}, Z) \\ \text{Station}(S) &\leftarrow U\text{-Bahn-Netz}(L, S, Z) \\ \text{Station}(Z) &\leftarrow U\text{-Bahn-Netz}(L, S, Z) \\ \text{Ans}(Z) &\leftarrow \text{Station}(Z), \neg \text{Erreichbar_BW}(Z) \end{aligned}$$

Hier: Semantik intuitiv klar.

Folie 169

Probleme mit der Semantik von Datalog[¬]

Zur Erinnerung: In der Übung wurde gezeigt, dass für jedes Datalog-Programm P und jede Datenbank $I \in \text{inst}(\text{edb}(P))$ gilt: für jedes $i \in \mathbb{N}$ ist $S_P^i(\hat{\mathbf{I}}) = T_P^i(\hat{\mathbf{I}})$.

Also ist $\llbracket P \rrbracket(\mathbf{I}) = S_P^\omega(\hat{\mathbf{I}}) = T_P^\omega(\hat{\mathbf{I}})$.

Für Datalog[¬] gilt dies nicht:

Beispiel 4.24. (a) $R(x) \leftarrow A(x), \neg R(x)$

“Ausgewertet” über einer DB $\mathbf{I} \in \text{inst}(\{A\})$ gilt:

$$S_P^i(\hat{\mathbf{I}})(R) = \mathbf{I}(A) \quad \text{für alle } i \geq 1$$

aber

$$T_P^i(\hat{\mathbf{I}})(R) = \begin{cases} \emptyset & \text{falls } i \text{ gerade} \\ \text{adom}(\mathbf{I}) & \text{falls } i \text{ ungerade} \end{cases}$$

Somit: Die Folge $(T_P^i(\hat{\mathbf{I}}))_{i \geq 0}$ hat keinen Fixpunkt.

Außerdem: $T_P(\cdot)$ hat überhaupt keinen Fixpunkt.

Folie 170

(b) $R(x) \leftarrow A(x), \neg S(x)$
 $S(x) \leftarrow A(x), \neg R(x)$

Hier gilt:

$$S_P^i(\hat{\mathbf{I}})(R) = S_P^i(\hat{\mathbf{I}})(S) = \mathbf{I}(A) \quad \text{für alle } i \geq 1$$

aber $T_P(\cdot)$ hat zwei verschiedene minimale Fixpunkte (*minimal bzgl. \subseteq*):

- FP₁ mit FP₁(R) = \emptyset und FP₁(S) = adom(**I**)
- FP₂ mit FP₂(R) = adom(**I**) und FP₂(S) = \emptyset

Folie 171

Datalog[⊃] und der Stufenoperator S_P

Um eine Semantik für Datalog[⊃] festzulegen, könnte man einfach wieder den Stufenoperator S_P betrachten. Natürlich gilt für jedes Datalog[⊃]-Programm P und jede Datenbank $\mathbf{J} \in \text{inst}(\text{sch}(P))$, dass

$$\mathbf{J} \subseteq S_P(\mathbf{J}) \subseteq S_P(S_P(\mathbf{J})) \subseteq \dots \subseteq S_P^i(\mathbf{J}) \subseteq S_P^{i+1}(\mathbf{J}) \subseteq \dots$$

Da $\text{adom}(P, \mathbf{J})$ endlich ist, wird irgendwann ein (eindeutig definierter) Fixpunkt von $S_P(\cdot)$ erreicht (dieser heißt übrigens *inflationärer Fixpunkt von P auf \mathbf{J}* , kurz: $S_P^\omega(\mathbf{J})$).

Wir könnten nun einfach festlegen, dass die Semantik von P auf $\mathbf{I} \in \text{inst}(\text{edb}(P))$ via $\llbracket P \rrbracket(\mathbf{I}) = S_P^\omega(\hat{\mathbf{I}})$ definiert ist.

Folie 172

Problem mit der inflationären Fixpunkt-Semantik von Datalog[⊃]

Diese über den Stufenoperator S_P definierte Semantik ist für viele Datalog[⊃]-Programme *unnatürlich*.

Beispiel 4.25.

$$\begin{aligned} E(L, S, Z) &\leftarrow U\text{-Bahn-Netz}(L, S, Z) \\ E(L, S, Z) &\leftarrow E(L, S, Y), U\text{-Bahn-Netz}(L, Y, Z) \\ \text{Erreichbar_BW}(Z) &\leftarrow E(L, \text{“Bockenheimer Warte”}, Z) \\ \text{Station}(S) &\leftarrow U\text{-Bahn-Netz}(L, S, Z) \\ \text{Station}(Z) &\leftarrow U\text{-Bahn-Netz}(L, S, Z) \\ \text{Ans}(Z) &\leftarrow \text{Station}(Z), \neg \text{Erreichbar_BW}(Z) \end{aligned}$$

“Intuitive Semantik”:

Alle Stationen, die von “Bockenheimer Warte” aus nicht ohne Umsteigen zu erreichen sind.

Inflationäre Fixpunkt-Semantik: Alle Stationen.

Folie 173

Frage: Was soll die Semantik von Datalog⁻ sein?

Probleme:

- Inflationäre Fixpunkt-Semantik: unnatürlich (siehe Beispiel 4.25)
- Fixpunkt-Semantik via $T_P^\omega(\hat{\mathbf{I}})$:
für manche Datalog⁻-Programme undefiniert (siehe Beispiel 4.24)
- Semantik via “kleinster Fixpunkt” von $T_P(\cdot)$:
ist für manche Datalog⁻-Programme nicht eindeutig (siehe
Beispiel 4.24)
- Beweisbasierte Semantik für Datalog⁻: ???

Aber für Programme wie in Beispiel 4.25 ist die Semantik “intuitiv klar”.

↪ Betrachte im Folgenden nur Datalog⁻-Programme von eingeschränkter Form:

- Semipositives Datalog⁻
- nr-Datalog⁻
- Stratifiziertes Datalog⁻

Folie 174

Semipositives Datalog⁻

Negation ist nur bei edb-Prädikaten erlaubt.

Man kann leicht zeigen, dass für jedes semipositive Datalog⁻-Programm P und

alle $\mathbf{I} \in \text{inst}(\text{edb}(P))$ gilt:

- $T_P(\cdot)$ hat einen eindeutig bestimmten kleinsten Fixpunkt \mathbf{J} mit $\mathbf{J}|_{\text{edb}(P)} = \mathbf{I}$.
- Dieser wird von der Sequenz

$$\hat{\mathbf{I}}, T_P(\hat{\mathbf{I}}), T_P^2(\hat{\mathbf{I}}), T_P^3(\hat{\mathbf{I}}), \dots$$

erreicht. Notation für diesen Fixpunkt: $T_P^\omega(\hat{\mathbf{I}})$.

Definition der Semantik von semipositiven Datalog[¬]-Programmen P :
 Für alle $\mathbf{I} \in \text{inst}(\text{edb}(P))$ setze

$$\llbracket P \rrbracket(\mathbf{I}) := T_P^\omega(\hat{\mathbf{I}})$$

Folie 175

Stratifiziertes Datalog[¬] — Beispiel

$$\begin{aligned} E(L, S, Z) &\leftarrow U\text{-Bahn-Netz}(L, S, Z) \\ E(L, S, Z) &\leftarrow E(L, S, Y), U\text{-Bahn-Netz}(L, Y, Z) \\ \text{Erreichbar_BW}(Z) &\leftarrow E(L, \text{“Bockenheimer Warte”}, Z) \\ \text{Station}(S) &\leftarrow U\text{-Bahn-Netz}(L, S, Z) \\ \text{Station}(Z) &\leftarrow U\text{-Bahn-Netz}(L, S, Z) \\ \text{Ans}(Z) &\leftarrow \text{Station}(Z), \neg \text{Erreichbar_BW}(Z) \end{aligned}$$

- Die Negation ist hier nicht mit der Rekursion verschränkt.
- Sie kann angewendet werden, nachdem $\text{Erreichbar_BW}(\cdot)$ vollständig berechnet ist.
- \rightsquigarrow Grundidee für stratifiziertes Datalog[¬].

Folie 176

Stratifiziertes Datalog[¬] — Präzise

Definition 4.26. Sei P ein Datalog[¬]-Programm.

Eine *Stratifizierung* von P ist eine Folge P^1, \dots, P^m von Datalog[¬]-Programmen, so dass $m \geq 1$ ist und es eine Abbildung $\sigma : \text{idb}(P) \rightarrow \{1, \dots, m\}$ gibt, so dass gilt:

- (1) P^1, \dots, P^m ist eine *Partition* von P (d.h. $P = P^1 \dot{\cup} \dots \dot{\cup} P^m$),
- (2) für jedes idb-Prädikat R von P gilt: alle Regeln, in deren Kopf R vorkommt, gehören zu $P^{\sigma(R)}$,
- (3) kommt ein $S \in \text{idb}(P)$ im Rumpf einer Regel mit Kopf R vor, so ist $\sigma(S) \leq \sigma(R)$, und

(4) kommt ein $S \in \text{idb}(P)$ *negiert* im Rumpf einer Regel mit Kopf R vor, so ist $\sigma(S) < \sigma(R)$.

Notation:

- P^i heißt *i-tes Stratum* bzw. *i-te Schicht* der Stratifizierung P^1, \dots, P^m
(“Stratum”: lat. für “Schicht”; Plural: Strata)
- σ heißt *Stratifizierungs-Abbildung*
- Ein Datalog⁻-Programm P heißt *stratifizierbar*, falls es eine Stratifizierung von P gibt. *Stratifiziertes Datalog⁻* bezeichnet die Menge aller stratifizierbaren Datalog⁻-Programme.

Folie 177

Stratifiziertes Datalog⁻ — Beispiele

$P^1 :=$

$$\begin{aligned} E(L, S, Z) &\leftarrow BVG(L, S, Z) \\ E(L, S, Z) &\leftarrow E(L, S, Y), U\text{-Bahn-Netz}(L, Y, Z) \\ Erreichbar_BW(Z) &\leftarrow E(L, \text{“Bockenheimer Warte”}, Z) \\ Station(S) &\leftarrow U\text{-Bahn-Netz}(L, S, Z) \\ Station(Z) &\leftarrow U\text{-Bahn-Netz}(L, S, Z) \end{aligned}$$

$P^2 :=$

$$Ans(Z) \leftarrow Station(Z), \neg Erreichbar_BW(Z)$$

ist eine Stratifizierung des Datalog⁻-Programms aus Beispiel 4.25.

Das Datalog⁻-Programm $R(x) \leftarrow A(x), \neg R(x)$ ist *nicht stratifizierbar*.

Das Datalog⁻-Programm

$$\begin{aligned} R(x) &\leftarrow A(x), \neg S(x) \\ S(x) &\leftarrow A(x), \neg R(x) \end{aligned}$$

auch nicht.

Folie 178

Test auf Stratifizierbarkeit

Definition 4.27. Sei P ein Datalog⁻-Programm. Der *Abhängigkeitsgraph* $G_P = (V_P, E_P^+, E_P^-)$ ist der gerichtete Graph mit

- Knotenmenge $V_P := sch(P)$
- Kantenmengen

$$E_P^+ := \left\{ (R, S) : \begin{array}{l} \text{es gibt eine Regel in } P, \text{ in deren Kopf} \\ S \text{ vorkommt und in deren Rumpf } R \text{ po-} \\ \text{sitiv vorkommt} \end{array} \right\}$$

$$E_P^- := \left\{ (R, S) : \begin{array}{l} \text{es gibt eine Regel in } P, \text{ in deren Kopf} \\ S \text{ vorkommt und in deren Rumpf } R \text{ ne-} \\ \text{gativ vorkommt} \end{array} \right\}$$

Beispiel: Siehe Tafel: Abhängigkeitsgraph für

$$\begin{aligned} E(L, S, Z) &\leftarrow U\text{-Bahn-Netz}(L, S, Z) \\ E(L, S, Z) &\leftarrow E(L, S, Y), U\text{-Bahn-Netz}(L, Y, Z) \\ Erreichbar_BW(Z) &\leftarrow E(L, \text{“Bockenheimer Warte”}, Z) \\ Station(S) &\leftarrow U\text{-Bahn-Netz}(L, S, Z) \\ Station(Z) &\leftarrow U\text{-Bahn-Netz}(L, S, Z) \\ Ans(Z) &\leftarrow Station(Z), \neg Erreichbar_BW(Z) \end{aligned}$$

Folie 179

Test auf Stratifizierbarkeit

Proposition 4.28. Für jedes Datalog⁻-Programm P gilt:

$$P \text{ ist stratifizierbar} \iff \begin{array}{l} \text{Im Abhängigkeitsgraph } G_P \text{ gibt es} \\ \text{keinen Kreis, in dem eine Kante} \\ \text{aus } E_P^- \text{ vorkommt.} \end{array}$$

Beweis:

“ \implies ”: Sei σ eine Stratifizierungs-Abbildung von P .

Angenommen, es gibt einen Kreis von R nach R , auf dem mindestens eine Kante aus E_P^- vorkommt.

Dann gilt: $\sigma(R) > \sigma(R)$. *Widerspruch!*

“ \impliedby ”: Idee: Nutze eine *topologische Sortierung* der *starken Zusammenhangskomponenten* von $(E_P^+ \cup E_P^-)$, um die einzelnen Schichten zu definieren.

Details: *Übung*.

Folie 180

Stratifiziertes Datalog[¬] — Semantik

- Sei P^1, \dots, P^m eine Stratifizierung eines Datalog[¬]-Programms P .
- Betrachte jede Schicht P^i als ein semipositives Datalog[¬]-Programm mit
 $edb(P^i) \subseteq edb(P) \cup idb(P^1) \cup \dots \cup idb(P^{i-1})$
- Sei $\mathbf{I} \in edb(P)$.

Die *Semantik* $\llbracket P \rrbracket(\mathbf{I})$ von P auf \mathbf{I} ist folgendermaßen definiert:

$$\llbracket P \rrbracket(\mathbf{I}) := \mathbf{I}^m,$$

wobei

$$\begin{aligned} \mathbf{I}^1 &:= \llbracket P^1 \rrbracket(\mathbf{I}) \\ \mathbf{I}^2 &:= \llbracket P^2 \rrbracket(\mathbf{I}^1) \\ &\vdots \\ \mathbf{I}^m &:= \llbracket P^m \rrbracket(\mathbf{I}^{m-1}) \end{aligned}$$

Man kann leicht zeigen, dass Folgendes gilt:

- (a) Obige Definition hängt nicht von der konkreten Wahl der Stratifizierung von P ab.
 D.h. für je zwei verschiedene Stratifizierungen P^1, \dots, P^m und Q^1, \dots, Q^n von P gilt: $\llbracket P^m \rrbracket(\mathbf{I}^{m-1}) = \llbracket Q^n \rrbracket(\mathbf{I}^{n-1})$.
- (b) $\llbracket P \rrbracket(\mathbf{I})$ ist der (eindeutig definierte) kleinste Fixpunkt \mathbf{J} von $T_P(\cdot)$ mit $\mathbf{J}|_{edb(P)} = \mathbf{I}$.

Folie 181

Spezialfall: nr-Datalog[¬]

nr-Datalog[¬] = nr-Datalog mit Negation
 = stratifiziertes Datalog[¬], eingeschränkt auf Programme P , in deren Abhängigkeitsgraph es keinen gerichteten Kreis (über $(E_P^+ \cup E_P^-)$) gibt.

Satz 4.29. *Die folgenden Anfragesprachen können genau dieselben Anfragefunktionen beschreiben:*

(a) *nr-Datalog⁻-Anfragen*

(b) *Relationale Algebra*

(c) *Relationenkalkül.*

Bemerkung: Die Übersetzung von nr-Datalog⁻ in eine der anderen Sprachen kann mehr als polynomiell viel Zeit beanspruchen, da nr-Datalog⁻-Programme u.U. viel kürzer sind als äquivalente Anfragen der anderen Sprachen.

Beweis: Der Beweis wird in den folgenden Kapiteln zur Relationalen Algebra (Kapitel 6) und zum Relationenkalkül (Kapitel 7) gegeben.

Kapitel 5

Funktionale Abhängigkeiten

5.1 Notationen

Folie 182

Zur Erinnerung — Benannte Perspektive

In Kapitel 2 haben wir festgelegt:

- Eine abzählbar unendliche Menge **att** von *Attributnamen*. Diese Menge ist *geordnet* via $\leq_{\mathbf{att}}$.
- Eine abzählbar unendliche Menge **rel** von *Relationsnamen*. Die Mengen **att**, **dom**, **rel** sind disjunkt.
- Eine Funktion $\mathit{sorte} : \mathbf{rel} \rightarrow \mathcal{P}_e(\mathbf{att})$, die jedem Relationsnamen eine endliche Menge von Attributen zuordnet ... und zwar so, dass f.a. $U \subseteq_e \mathbf{att}$ gilt: es gibt unendlich viele $R \in \mathbf{rel}$ mit $\mathit{sorte}(R) = U$.
- Ein *Relationsschema* ist einfach ein Relationsname R .
- Manchmal schreiben wir kurz $R[U]$ für $\mathit{sorte}(R) = U$.
- Ein *R-Tupel* ist eine Funktion $t : \mathit{sorte}(R) \rightarrow \mathbf{dom}$.
- Eine *R-Relation* ist eine endliche Menge von R -Tupeln.
- $\mathit{inst}(R)$ bezeichnet die Menge aller R -Relationen.
- $\mathit{inst}(U)$ bezeichnet die Menge aller Relationen über einem Relationsschema der Sorte U (für $U \subseteq_e \mathbf{att}$)

Folie 183

[label=NFMotivation1]

Motivation

Ziel beim Datenbank-Entwurf:

Ein DB-Schema entwickeln, so dass Informationen zum gewünschten Anwendungsbereich „sinnvoll“ gespeichert werden können.
 Insbesondere: *Wenn möglich, Redundanzen und Inkonsistenzen vermeiden.*

Beispiel: Relation *Warenlager*[Bauteil-Nr, Lager-Nr, Menge, Ort]

Warenlager:

Bauteil-Nr	Lager-Nr	Menge	Ort
2411	2	200	Riedberg
2412	3	300	Bornheim
3001	1	100	Hanau
2415	2	100	Riedberg

Unschön: Redundanz — der Ort von Lager 2 ist mehrfach gespeichert.
 Dadurch können Inkonsistenzen auftreten: *Update-Anomalien = Inkonsistenzen, die durch Aktualisierung der DB auftreten können:*

- *Änderungs-Anomalie:* den Ort in Zeile 1 durch “Westend” ersetzen
 ~> Adresse von Lager 2 nicht mehr eindeutig
- *Lösch-Anomalie:* Löschen von Zeile 2
 ~> Information über die Adresse von Lager 3 geht verloren
- *Einfüge-Anomalie:* Die Adresse eines neuen Lagers kann erst dann eingefügt werden, wenn mindestens ein Bauteil dort gelagert wird.

Folie 184

[label=NFMotivation2] **Zur Vermeidung dieser Update-Anomalien:**

Informationen auf 2 Relationen *Adressen*[Lager-Nr, Ort] und *Lagerung*[Bauteil-Nr, Lager-Nr, Menge] aufteilen

Adressen:

Lager-Nr	Ort
2	Riedberg
3	Bornheim
1	Hanau

Abhängigkeit:

Lager-Nr → Ort

Lagerung:

Bauteil-Nr	Lager-Nr	Menge
2411	2	200
2412	3	300
3001	1	100
2415	2	100

Abhängigkeit:

Bauteil-Nr, Lager-Nr \rightarrow Menge **Zur Anfrage-Optimierung:**

Die „optimierte Anfrage“ muss nur auf solchen Datenbanken äquivalent zur Original-Anfrage sein, die die obigen Abhängigkeiten erfüllen.

\leadsto die minimale, solchermaßen äquivalente Anfrage ist evtl. noch kleiner als die, die durch die Tableau-Minimierung aus Kapitel 3.4 gefunden wird.

Folie 185

Notation

Attributnamen : A, B, C, A_1, A_2, \dots

Attributmengen : $U, X, Y, Z, X_1, X_2, \dots$ (endliche Mengen von Attributnamen)

Relationen : I, J

$\{A, B, C\}$: ABC

$X \cup Y$: XY

Folie 186

Funktionale Abhängigkeiten

Definition 5.1. Sei U eine endliche Menge von Attributnamen.

- (a) Eine *funktionale Abhängigkeit* (kurz: *FD*) über U ist ein Ausdruck der Form $X \rightarrow Y$, wobei $X, Y \subseteq U$. („FD“ steht für „functional dependency“)
- (b) Eine Relation $I \in \text{inst}(U)$ erfüllt die FD $X \rightarrow Y$ (kurz: $I \models X \rightarrow Y$), falls für alle Tupel t und s aus I gilt:

$$\pi_X(t) = \pi_X(s) \implies \pi_Y(t) = \pi_Y(s)$$

D.h.: Wenn t und s in sämtlichen Spalten aus X übereinstimmen, dann stimmen sie auch in jeder Spalte aus Y überein.

(c) Ist \mathcal{F} eine Menge von FDs über U , so gilt

$$I \models \mathcal{F} \quad : \iff \quad \text{für alle } f \in \mathcal{F} \text{ gilt } I \models f.$$

(d) Eine *Schlüsselbedingung* ist eine FD der Form $X \rightarrow U$.

Folie 187

Funktionale Abhängigkeiten und verlustfreie Joins

Proposition 5.2. Sei $X \rightarrow Y$ eine FD über U und sei $Z := U \setminus (X \cup Y)$.

Für jede Relation $I \in \text{inst}(U)$ gilt:

Falls $I \models X \rightarrow Y$, so ist $I = \pi_{XY}(I) \bowtie \pi_{XZ}(I)$.

Beweis: Siehe Tafel.

Folgerung: Die in Relation I gespeicherte Information kann „verlustfrei“ auf zwei Relationen aufgeteilt werden (eine mit den Spalten XY und eine mit den Spalten XZ), aus denen die Original-Relation rekonstruiert werden kann. (*Stichwort:* „lossless join“)

Beispiel für einen „verlustreichen Join“: Siehe Tafel.

5.2 The Chase — Die Verfolgungsjagd

Folie 188

Beispiel zu „The Chase — Die Verfolgungsjagd“

Beispiel 5.3. Tableau-Anfrage $Q = (T, t)$ mit

$$T = \begin{array}{|c|c|c|c|} \hline A & B & C & D \\ \hline w & x & y & z' \\ \hline w' & x & y' & z \\ \hline \end{array} \quad \text{und} \quad t = (w, x, y, z)$$

Klar: Die Anfrage $Q := (T, t)$ ist *minimal* im Sinne von Kapitel 3.4.

Situation jetzt:

- Gegeben sei die FD-Menge $\mathcal{F} := \{ B \rightarrow D \}$
- Q soll nur auf solchen DBs ausgewertet werden, die \mathcal{F} erfüllen
- Ziel: Vereinfache (minimiere) Q .

Details: siehe Tafel.

Folie 189

Äquivalenz bzw. Query Containment bzgl. \mathcal{F}

Für ein Relationsschema R und das Datenbankschema $\mathbf{S} := \{R\}$

- identifizieren wir eine R -Relation I mit der Datenbank $\mathbf{I} \in \text{inst}(\mathbf{S})$ mit $\mathbf{I}(R) = I$.
- sagen wir auch „ Q ist eine Anfrage über R “ statt „ Q ist eine Anfrage über \mathbf{S} “.
- schreiben wir $I \models Q$ an Stelle von $\mathbf{I} \models Q$, für eine R -Relation I und die zugehörige Datenbank \mathbf{I} mit $\mathbf{I}(R) = I$. Analog schreiben wir $\llbracket Q \rrbracket(I)$ statt $\llbracket Q \rrbracket(\mathbf{I})$.

Definition 5.4. Sei R ein Relationsschema, sei $\mathcal{I} \subseteq \text{inst}(R)$, und seien Q_1 und Q_2 Anfragen über R .

- $Q_1 \sqsubseteq_{\mathcal{I}} Q_2 \quad : \iff \quad$ für alle $I \in \mathcal{I}$ ist $\llbracket Q_1 \rrbracket(I) \subseteq \llbracket Q_2 \rrbracket(I)$
- $Q_1 \equiv_{\mathcal{I}} Q_2 \quad : \iff \quad$ für alle $I \in \mathcal{I}$ ist $\llbracket Q_1 \rrbracket(I) = \llbracket Q_2 \rrbracket(I)$.

Sei \mathcal{F} eine Menge von FDs über R .

- $\text{Mod}(\mathcal{F}) := \text{Mod}_R(\mathcal{F}) := \{ I \in \text{inst}(R) : I \models \mathcal{F} \}$
- $Q_1 \sqsubseteq_{\mathcal{F}} Q_2 \quad : \iff \quad Q_1 \sqsubseteq_{\text{Mod}(\mathcal{F})} Q_2$
- $Q_1 \equiv_{\mathcal{F}} Q_2 \quad : \iff \quad Q_1 \equiv_{\text{Mod}(\mathcal{F})} Q_2$

Folie 190

Vereinbarungen für den Rest von Kapitel 5.2

Der Einfachheit halber betrachten wir im Folgenden

- ein festes Relationsschema R
- Mengen \mathcal{F} von FDs über R , in denen o.B.d.A. jede FD von der Form $X \rightarrow A$ mit $X \subseteq \text{sorte}(R)$ und $A \in \text{sorte}(R)$ ist
- eine feste lineare Ordnung $<$ auf der Variablenmenge var
- nur Tableau-Anfragen $Q = (T, t)$ über R , in denen *keine Konstanten vorkommen*

Bemerkung: Die Ergebnisse aus Kapitel 5.2 können leicht verallgemeinert werden auf Anfragen mit Konstanten und auf Anfragen über einem Datenbankschema.

Folie 191

Regel für die Verfolgungsjagd

Definition 5.5. *FD-Regel:*

Sei $f := (X \rightarrow A)$ eine FD über R , sei (T, t) eine Tableau-Anfrage über R .
 Seien u und v Zeilen von T mit $\pi_X(u) = \pi_X(v)$ und $u(A) \neq v(A)$.
 Sei $\{x, y\} := \{u(A), v(A)\} \subseteq \text{var}$ und sei $x < y$.

Anwenden der FD f auf u, v in (T, t) liefert die Tableau-Anfrage $(h(T), h(t))$, wobei h die Substitution mit $h(y) := x$ und $h(z) := z$ für alle $z \in \text{Var}((T, t))$ mit $z \neq y$.

Folie 192

Anwenden der FD-Regel erhält Äquivalenz bzgl. \mathcal{F}

Proposition 5.6. *Sei \mathcal{F} eine Menge von FDs über R ,
 sei $f := (X \rightarrow A) \in \mathcal{F}$,
 sei $Q := (T, t)$ eine Tableau-Anfrage über R und
 sei $Q' := (T', t')$ eine Tableau-Anfrage, die durch 1-maliges Anwenden der
 FD-Regel mit einer FD $f \in \mathcal{F}$ aus Q entsteht.*

Dann gilt: $Q \equiv_{\mathcal{F}} Q'$.

Beweis: Siehe Tafel.

Folie 193

Verfolgungssequenzen

Definition 5.7. (a) Eine *Verfolgungssequenz* für (T, t) mittels \mathcal{F} ist eine Folge

$$(T_0, t_0), (T_1, t_1), (T_2, t_2), \dots$$

für die gilt:

- $(T_0, t_0) = (T, t)$ und
- für jedes $i \geq 0$ entsteht (T_{i+1}, t_{i+1}) durch 1-maliges Anwenden der FD-Regel mit einer FD aus \mathcal{F} auf (T_i, t_i) .

(b) Die Verfolgungssequenz ist *terminiert*, falls sie endlich ist und auf ihr letztes Element (T_m, t_m) keine FD-Regel mit einer FD aus \mathcal{F} mehr angewendet werden kann.
 (T_m, t_m) heißt dann das *Resultat* der Sequenz.

Notation: $T \models \mathcal{F}$

Definition 5.8. (a) Ein Tableau T über R erfüllt die FD $X \rightarrow A$ (kurz: $T \models X \rightarrow A$), falls
für alle Zeilen u und v von T gilt:

$$\pi_X(u) = \pi_X(v) \implies u(A) = v(A)$$

(D.h.: Wenn u und v in sämtlichen Spalten aus X übereinstimmen, dann stimmen sie auch in der Spalte A überein.)

(b) Ist \mathcal{F} eine Menge von FDs über R , so ist

$$T \models \mathcal{F} \iff T \models f, \text{ für alle } f \in \mathcal{F}.$$

Eigenschaften des Resultats einer Verfolgungssequenz

Lemma 5.9. Sei (T', t') das Resultat einer terminierten Verfolgungssequenz für (T, t) mittels \mathcal{F} .
Dann gilt: $(T', t') \equiv_{\mathcal{F}} (T, t)$ und $T' \models \mathcal{F}$.

Beweis: Übung.

Beobachtung: Jede Verfolgungssequenz ist endlich und kann zu einer terminierten Sequenz vervollständigt werden.

Bemerkenswert: Man kann zeigen, dass alle terminierten Verfolgungssequenzen für (T, t) mittels \mathcal{F} dasselbe Resultat liefern. Dies wird auch *Church-Rosser-Eigenschaft* genannt.

Die Church-Rosser-Eigenschaft der Verfolgungsjagd

Theorem 5.10. Sei (T, t) eine Tableau-Anfrage über R und sei \mathcal{F} eine Menge von FDs über R . Dann gilt: Alle terminierten Verfolgungssequenzen für (T, t) mittels \mathcal{F} liefern dasselbe Resultat.

Hier ohne Beweis.

Ein Beweis findet sich am Ende von Kapitel 8.4 des Buchs [AHV].

Definition 5.11. Ist (T, t) eine Tableau-Anfrage über R und \mathcal{F} eine Menge von FDs über R , so bezeichnet $\text{chase}(T, t, \mathcal{F})$ das Resultat einer (bzw. sämtlicher) terminierter Verfolgungssequenzen für (T, t) mittels \mathcal{F} .

Bemerkung: Von Lemma 5.9 wissen wir, dass $\text{chase}(T, t, \mathcal{F}) \equiv_{\mathcal{F}} (T, t)$ und $\text{chase}(T, t, \mathcal{F}) \models \mathcal{F}$.

Folie 197

Berechnung von $\text{chase}(T, t, \mathcal{F})$

Korollar 5.12. *Es gibt einen Polynomialzeit-Algorithmus, der bei Eingabe einer Tableau-Anfrage (T, t) über R und einer Menge \mathcal{F} von FDs über R die Tableau-Anfrage $\text{chase}(T, t, \mathcal{F})$ berechnet.*

Beweis:

Algorithmus:

- (1) Wiederhole so lange, bis keine FD-Regel bzgl. \mathcal{F} mehr auf (T, t) anwendbar ist:
 - (1.1) (T', t') sei das Resultat der Anwendung einer FD-Regel bzgl. \mathcal{F} auf (T, t) .
 - (1.2) Setze $(T, t) := (T', t')$.
- (2) Gib (T, t) aus.

Korrektheit folgt direkt aus der Church-Rosser-Eigenschaft.

Polynomielle Laufzeit, da jede FD $f \in \mathcal{F}$ auf jedes Paar u, v von Zeilen von T höchstens 1-mal angewendet werden kann und da jede einzelne Anwendung der FD-Regel nur polynomiell viel Zeit benötigt. □

Folie 198

Äquivalenz von Anfragen bzgl. \mathcal{F}

Theorem 5.13. Seien $Q_1 := (T_1, t_1)$ und $Q_2 := (T_2, t_2)$ Tableau-Anfragen über R und sei \mathcal{F} eine Menge von FDs über R . Dann gilt:

$$(a) \quad Q_1 \sqsubseteq_{\mathcal{F}} Q_2 \iff \text{chase}(T_1, t_1, \mathcal{F}) \sqsubseteq \text{chase}(T_2, t_2, \mathcal{F}) \quad \text{und}$$

$$(b) \quad Q_1 \equiv_{\mathcal{F}} Q_2 \iff \text{chase}(T_1, t_1, \mathcal{F}) \equiv \text{chase}(T_2, t_2, \mathcal{F}) \quad \text{und}$$

Beweis: Siehe Tafel.

Bemerkung: Aus Theorem 5.13, Korollar 5.12 und Korollar 3.36 folgt insbesondere, dass „ $Q_1 \sqsubseteq_{\mathcal{F}} Q_2$ “ bzw. „ $Q_1 \equiv_{\mathcal{F}} Q_2$ “ entscheidbar ist und zur Komplexitätsklasse NP gehört.

Folie 199

Anfrage-Minimierung bzgl. \mathcal{F}

Vorgehensweise:

- *Eingabe:* Tableau-Anfrage $Q = (T, t)$ über R und Menge \mathcal{F} von FDs über R
- *Schritt 1:* Berechne $Q' := (T', t') := \text{chase}(T, t, \mathcal{F})$. Klar: $Q' \equiv_{\mathcal{F}} Q$
- *Schritt 2:* Nutze den Algorithmus aus Theorem 3.39(a) um eine *minimale* Tableau-Anfrage $Q'' := (T'', t'')$ mit $Q'' \equiv Q'$ zu berechnen
Insbesondere gilt: $Q'' \equiv_{\mathcal{F}} Q' \equiv_{\mathcal{F}} Q$

Notation: Ist $Q = (T, t)$ eine Tableau-Anfrage, so schreibe $\text{min}(Q) := \text{min}(T, t)$, um die gemäß Theorem 3.39(a) minimale zu Q äquivalente Tableau-Anfrage zu bezeichnen.

Lemma 5.14. Sei $Q = (T, t)$ eine Tableau-Anfrage über R und sei \mathcal{F} eine Menge von FDs über R .

Dann gilt: $|\text{min}(\text{chase}(T, t, \mathcal{F}))| \leq |\text{min}(T, t)|$.

Hier ohne Beweis.

Eine Beweisskizze findet sich auf Seite 178 des Buchs [AHV].

5.3 Der Armstrong-Kalkül

Folie 200

Implikation von Abhängigkeiten

Gegeben: Eine Menge \mathcal{F} von FDs über U .

Frage: Welche anderen FDs folgen aus \mathcal{F} ?

Beispiel 5.15. Sei $U := \{A, B, C, D, E\}$ und $\mathcal{F} := \{A \rightarrow C, B \rightarrow C, CD \rightarrow E\}$.

Dann gilt für jede Relation I mit $I \models \mathcal{F}$ auch, dass $I \models AD \rightarrow E$.

Definition 5.16. Seien \mathcal{F} und \mathcal{G} zwei Mengen von FDs über U .

(a) \mathcal{F} impliziert \mathcal{G} (kurz: $\mathcal{F} \models_U \mathcal{G}$ bzw. $\mathcal{F} \models \mathcal{G}$, falls U aus dem Kontext klar ist), falls für alle Relationen $I \in \text{inst}(U)$ gilt: Falls $I \models \mathcal{F}$, so auch $I \models \mathcal{G}$.

Besteht \mathcal{G} aus einer einzigen FD f , so schreibe auch $\mathcal{F} \models f$ statt $\mathcal{F} \models \{f\}$.

(b) \mathcal{F} und \mathcal{G} sind äquivalent (kurz: $\mathcal{F} \equiv_U \mathcal{G}$ bzw. $\mathcal{F} \equiv \mathcal{G}$, falls U aus dem Kontext klar ist), falls $\mathcal{F} \models \mathcal{G}$ und $\mathcal{G} \models \mathcal{F}$.

(c) Die Hülle von \mathcal{F} über U (kurz: $\mathcal{F}^{*,U}$ bzw. \mathcal{F}^* , falls U aus dem Kontext klar ist), ist definiert als

$\mathcal{F}^{*,U} := \{X \rightarrow Y : X, Y \subseteq U \text{ und } \mathcal{F} \models X \rightarrow Y\}$.

Klar: Für alle $X \subseteq U$ und alle $Y \subseteq X$ gilt $X \rightarrow Y \in \mathcal{F}^{*,U}$.
Insbesondere: $2^{|U|} \leq |\mathcal{F}^{*,U}| \leq 2^{2 \cdot |U|}$.

Folie 201

Fragen: Wie kann man (1) testen, ob $\mathcal{F} \models X \rightarrow Y$?
(2) die Hülle $\mathcal{F}^{*,U}$ berechnen?

Problem: Die Definition von $\mathcal{F} \models X \rightarrow Y$ spricht über *alle* Relationen $I \in \text{inst}(U)$. Das sind unendlich viele (da **dom** unendlich ist).

Lösung für (1): Reduziere das Problem „ $\mathcal{F} \models X \rightarrow Y$?“ auf das Problem „ $Q_X \sqsubseteq_{\mathcal{F}} Q_Y$?“ für geeignete Tableau-Anfragen Q_X und Q_Y .

Beachte: Einen Algorithmus zum Lösen des letzteren Problems haben wir bereits kennengelernt.

Satz 5.17.

Es gibt einen Linearzeit-Algorithmus, der bei Eingabe einer endlichen Attributmengens U , einer Menge \mathcal{F} von FDs über U und einer FD $X \rightarrow Y$ über U zwei azyklische Tableau-Anfragen Q_X und Q_Y berechnet, für die gilt:
 $\mathcal{F} \models X \rightarrow Y \iff Q_X \sqsubseteq_{\mathcal{F}} Q_Y$.

Ob $Q_X \sqsubseteq_{\mathcal{F}} Q_Y$ gilt, kann in Zeit polynomiell in der Größe von U , \mathcal{F} , X und Y entschieden werden.

Beweis: Siehe Tafel.

Folie 202

Fragen: Wie kann man (1) testen, ob $\mathcal{F} \models X \rightarrow Y$?
 (2) die Hülle $\mathcal{F}^{*,U}$ berechnen?

Problem: Die Definition von $\mathcal{F} \models X \rightarrow Y$ spricht über *alle* Relationen $I \in \text{inst}(U)$. Das sind unendlich viele (da **dom** unendlich ist).

Lösung für (2): Finde eine endliche Regelmengens \mathcal{R} , mit deren Hilfe aus \mathcal{F} neue Abhängigkeiten hergeleitet werden können.

(Analog zum Begriff der „Kalküle“ aus der Vorlesung *Logik in der Informatik*)

Ziel:

- Alle FDs, die sich mit \mathcal{R} aus \mathcal{F} ableiten lassen, sind in $\mathcal{F}^* \rightsquigarrow \mathcal{R}$ *ist korrekt*
- Alle FDs aus \mathcal{F}^* lassen sich in \mathcal{R} aus \mathcal{F} herleiten $\rightsquigarrow \mathcal{R}$ *ist vollständig*

Folie 203

Der Armstrong-Kalkül

Der *Armstrong-Kalkül* \mathcal{R}_A über U ist wie folgt definiert:

Axiome:

(A) Für alle $X \subseteq U$ und alle $Y \subseteq X$ ist

$$\overline{X \rightarrow Y} \quad \text{ein Axiom des Armstrong-Kalküls über } U.$$

Weitere Regeln:

(E) Für alle $X, Y, Z \subseteq U$ ist die *Erweiterungsregel*

$$\frac{X \rightarrow Y}{XZ \rightarrow YZ} \quad \text{eine Regel des Armstrong-Kalküls über } U.$$

(T) Für alle $X, Y, Z \subseteq U$ ist die *Transitivitätsregel*

$$\frac{X \rightarrow Y \quad Y \rightarrow Z}{X \rightarrow Z} \quad \text{eine Regel des Armstrong-Kalküls über } U.$$

Folie 204

Ableitungen im Armstrong-Kalkül

Definition 5.18. Sei f eine FD über U und sei \mathcal{F} eine Menge von FDs über U .

Eine *Ableitung von f aus \mathcal{F} im Armstrong-Kalkül* ist eine endliche Folge (f_1, \dots, f_ℓ) von FDs über U , so dass $\ell \geq 1$ und

- $f_\ell = f$ und
- für alle $i \in \{1, \dots, \ell\}$ gilt:
 - $f_i \in \mathcal{F}$ oder
 - f_i ist ein Axiom (A) des Armstrong-Kalküls über U oder
 - es gibt ein $j < i$, so dass $\frac{f_j}{f_i}$ eine Erweiterungsregel (E) des Armstrong-Kalküls über U ist oder
 - es gibt $j, k < i$, so dass $\frac{f_j f_k}{f_i}$ eine Transitivitätsregel (T) des Armstrong-Kalküls über U ist.

Wir schreiben $\mathcal{F} \vdash_U f$ (bzw. $\mathcal{F} \vdash f$, falls U aus dem Kontext klar ist), um auszudrücken, dass es eine Ableitung von f aus \mathcal{F} im Armstrong-Kalkül über U gibt.

Mit $\text{abl}_{\mathcal{R}_A}(\mathcal{F})$ bezeichnen wir die Menge aller FDs f über U , für die gilt: $\mathcal{F} \vdash_U f$.

Folie 205

Beispiel für eine Ableitung im Armstrong-Kalkül

Sei $U := \{A, B, C, D, E\}$ und $\mathcal{F} := \{A \rightarrow C, B \rightarrow C, CD \rightarrow E\}$.

Eine Ableitung von $AD \rightarrow E$ aus \mathcal{F} im Armstrong-Kalkül über U :

$$\left(A \rightarrow C, AD \rightarrow CD, CD \rightarrow E, AD \rightarrow E \right)$$

Erläuterung:

- (1) $A \rightarrow C \in \mathcal{F}$
- (2) $AD \rightarrow CD$ durch Anwenden der Erweiterungsregel (E) auf (1)
- (3) $CD \rightarrow E \in \mathcal{F}$
- (4) $AD \rightarrow E$ durch Anwenden der Transitivitätsregel (T) auf (2) und (3).

Folie 206

Korrektheit und Vollständigkeit des Armstrong-Kalküls

Sei U eine endliche Menge von Attributnamen und sei \mathcal{F} eine Menge von FDs über U .

Der Armstrong-Kalkül \mathfrak{K}_A über U heißt

- *korrekt*, falls für jede Menge \mathcal{F} von FDs über U und jede FD f über U gilt:
Falls $\mathcal{F} \vdash_U f$, so $\mathcal{F} \models_U f$.
- *vollständig*, falls für jede Menge \mathcal{F} von FDs über U und jede FD f über U gilt:
Falls $\mathcal{F} \models_U f$, so $\mathcal{F} \vdash_U f$.

Theorem 5.19 (Armstrong, 1974). *Der Armstrong-Kalkül \mathfrak{K}_A ist korrekt und vollständig, d.h.:*

Für jede endliche Menge U von Attributnamen, jede FD f über U und jede Menge \mathcal{F} von FDs über U gilt: $\mathcal{F} \vdash_U f \iff \mathcal{F} \models_U f$.

Beweis: “ \implies ”: Übung. “ \impliedby ”: Siehe Tafel.

Bemerkungen:

- Die Hülle $\mathcal{F}^{*,U}$ kann man also berechnen, indem man alle FDs f berechnet, für die gilt $\mathcal{F} \vdash_U f$.

- Wenn man nur für eine bestimmte FD f testen will, ob $\mathcal{F} \models_U f$, so ist es ziemlich aufwändig, erst ganz $\mathcal{F}^{*,U}$ zu berechnen, da die Hülle immer mindestens $2^{|U|}$ viele Elemente enthält.

Kapitel 6

Relationale Algebra

6.1 Definition und Beispiele

Grenzen der Ausdrucksstärke konjunktiver Anfragen

Folie 207

Wir haben gesehen:

- konjunktive Anfragen können nur *monotone* Anfragefunktionen beschreiben (Satz 3.6) \rightsquigarrow Anfragen mit Negationen der Art

Welche Regisseure haben noch nie mit "Matt Damon" gearbeitet?

können nicht in SPC- bzw. SPJR-Algebra gestellt werden.

- konjunktive Anfragen können keine Ver-ODER-ungen der Art

In welchen Kinos wird "Alien" oder "Brazil" gespielt?

ausdrücken (vgl. Übungsblatt 1).

Jetzt:

Erweitere SPC- bzw. SPJR-Algebra um die Möglichkeit, auch solche Anfragen zu beschreiben.

Folie 208

Vereinigung und Differenz

Operatoren \cup und $-$:

Diese Operatoren können angewendet werden auf Relationen I und J , die dieselbe Sorte bzw. Stelligkeit haben und liefern als Ausgabe die Relationen

$$I \cup J := \{t : t \in I \text{ oder } t \in J\}$$

bzw.

$$I - J := \{t \in I : t \notin J\}$$

Folie 209

SPJRU, SPCU und relationale Algebra

Definition 6.1. Sei \mathbf{S} ein Datenbankschema.

(a) Zur Definition der Klasse der Anfragen der $\text{SPJRU}[\mathbf{S}]$ (bzw. der $\text{SPCU}[\mathbf{S}]$) werden die Definitionen von $\text{SPJR}[\mathbf{S}]$ (bzw. $\text{SPC}[\mathbf{S}]$) um die folgende Regel erweitert:

- Sind Q und P zwei $\text{SPJRU}[\mathbf{S}]$ -Anfragen derselben Sorte Σ (bzw. $\text{SPCU}[\mathbf{S}]$ -Anfragen derselben Stelligkeit k), so ist $(Q \cup P)$ eine $\text{SPJRU}[\mathbf{S}]$ -Anfrage der Sorte Σ (bzw. eine $\text{SPCU}[\mathbf{S}]$ -Anfrage der Stelligkeit k).

(b) Zur Definition der Klasse der Anfragen der *relationalen Algebra* über \mathbf{S} in der benannten (bzw. der unbenannten) Perspektive werden die Definitionen von $\text{SPJRU}[\mathbf{S}]$ (bzw. $\text{SPCU}[\mathbf{S}]$) um die folgende Regel erweitert:

- Sind Q und P zwei Anfragen der relationalen Algebra derselben Sorte Σ (bzw. derselben Stelligkeit k), so ist $(Q - P)$ eine Anfrage relationalen Algebra der Sorte Σ (bzw. der Stelligkeit k).

Die *Semantik* $\llbracket Q \rrbracket$ solcher Anfragen Q ist induktiv auf die offensichtliche Art definiert.

Mit SPJRU (bzw. SPCU) bezeichnen wir die Klasse aller $\text{SPJRU}[\mathbf{S}]$ -Anfragen (bzw. $\text{SPCU}[\mathbf{S}]$ -Anfragen) für alle Datenbankschemata \mathbf{S} .

Folie 210

Beispiele

- In welchen Kinos wird “Alien” oder “Brazil” gespielt?

$$\pi_{Kino} \left(\sigma_{Titel="Alien"}(Programm) \cup \sigma_{Titel="Brazil"}(Programm) \right)$$

- Welche Regisseure haben noch nie mit “Matt Damon” gearbeitet?

$$\pi_{Regie}(Filme) - \pi_{Regie} \left(\sigma_{Schauspieler="Matt Damon"}(Filme) \right)$$

- Welche derzeit laufenden Filme haben nur Schauspieler, die schon mal in einem Film von “James Cameron” mitgespielt haben?

$$\pi_{Titel}(Programm) - \pi_{Titel} \left(Filme \bowtie \underbrace{\left(\pi_{Schauspieler}(Filme) - \pi_{Schauspieler} \left(\sigma_{Regie="James Cameron"}(Filme) \right) \right)}_{\text{Schauspieler, die noch nie mit James Cameron gearbeitet haben}} \right)$$

Filme mit mind. einem Schauspieler, der noch nie mit James Cameron gearbeitet hat

Folie 211

Ausdrucksstärke

Proposition 6.2.

(a) Jede SPCU-Anfrage und jede SPJRU-Anfrage ist monoton.

(b) Für jede Datenbank \mathbf{I} und jede Anfrage Q der relationalen Algebra gilt:

$$\text{adom}(\llbracket Q \rrbracket(\mathbf{I})) \subseteq \text{adom}(Q, \mathbf{I}).$$

(c) $\text{SPC} < \text{SPCU} < \text{relationale Algebra (unbenannte Perspektive)}$
 $\equiv \quad \equiv \quad \equiv$
 $\text{SPJR} < \text{SPJRU} < \text{relationale Algebra (benannte Perspektive)}$

Beweis:

(a)+(b): Einfache Induktion über den Aufbau der Anfragen.

(c): Übung. □

Folie 212

Proposition 6.3. (a) *Benannte Perspektive:*

Keiner der Operatoren $\sigma, \pi, \cup, -, \bowtie, \delta$ ist redundant.

D.h.: Weglassen jedes einzelnen dieser Operatoren führt zu einer Algebra, die manche der in der relationalen Algebra ausdrückbaren Anfragefunktionen nicht beschreiben kann.

(b) *Unbenannte Perspektive:*

(i) *Der Operator σ kann durch Kombination der Operatoren $\pi, -, \times$ ausgedrückt werden.*

Beachte: Um dies zu zeigen, muss man nutzen, dass bei der Projektion π_{j_1, \dots, j_k} die Indizes j_i nicht paarweise verschieden sein müssen.

(ii) *Keiner der Operatoren $\pi, \cup, -, \times$ ist redundant.*

Beweis: Übung.

Folie 213

Theta-Join und Semijoin

Eine *positive konjunktive Join-Bedingung* ist ein Ausdruck θ der Form $\bigwedge_{\ell=1}^m x_{i_\ell} = y_{j_\ell}$, für natürliche Zahlen $m \geq 0$ und $i_1, \dots, i_m, j_1, \dots, j_m \geq 1$.

Zwei Tupel $\bar{a} = (a_1, \dots, a_r) \in \mathbf{dom}^r$ und $\bar{b} = (b_1, \dots, b_s) \in \mathbf{dom}^s$ mit $r \geq \max\{i_1, \dots, i_m\}$ und $s \geq \max\{j_1, \dots, j_m\}$ erfüllen θ

$$(\text{kurz: } (\bar{a}, \bar{b}) \models \theta, \text{ bzw. } \theta(\bar{a}, \bar{b})),$$

falls für alle $\ell \in \{1, \dots, m\}$ gilt: $a_{i_\ell} = b_{j_\ell}$.

In der relationalen Algebra (unbenannte Perspektive) lassen sich u.a. die folgenden Operationen ausdrücken:

- *Theta-Join* \bowtie_θ , wobei θ eine positive konjunktive Join-Bedingung ist.
Semantik: $I \bowtie_\theta J := \{ (\bar{a}, \bar{b}) : \bar{a} \in I, \bar{b} \in J, \text{ so dass } (\bar{a}, \bar{b}) \models \theta \}$
- *Semijoin* \ltimes_θ , wobei θ eine positive konjunktive Join-Bedingung ist.
Semantik: $I \ltimes_\theta J := \{ \bar{a} \in I : \text{ex. } \bar{b} \in J, \text{ so dass } (\bar{a}, \bar{b}) \models \theta \}$

6.2 Anfrageauswertung und Heuristische Optimierung

Folie 214

Anfrageauswertung und Heuristische Optimierung

... hat viele Aspekte:

- Speicher- und Indexstrukturen
- Betriebssystem
- Seitenersetzungsstrategien
- Statistische Eigenschaften der Daten
- Statistische Informationen über Anfragen
- Implementierung der einzelnen Operatoren
- Ausdrucksstärke der Anfragesprache

Hier: Überblick und einige Teilaspekte.

Details: In den Vorlesungen DBS I und DBS II von Prof. Freytag

Folie 215

Anfrageauswertung allgemein

Vorbemerkung:

- Datenbanken sind *sehr* groß
- werden auf Sekundärspeicher (Festplatte) gespeichert
- *Aufwand wird dominiert durch die Anzahl der Plattenzugriffe* (“Seitenzugriffe”)
Denn: In derselben Zeit, die für einen “Random Access” auf der Festplatte benötigt wird, können zigtausende Operationen im Hauptspeicher durchgeführt werden.

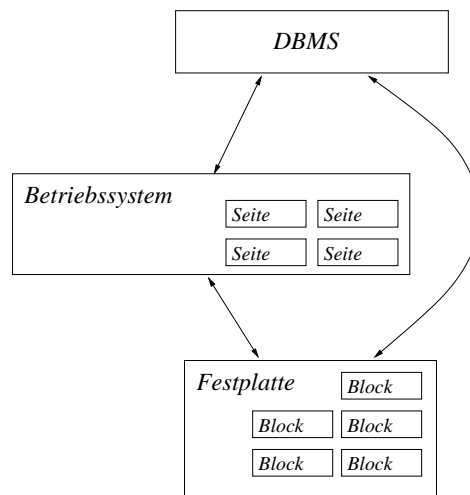
Allgemeines Vorgehen:

- durch Erzeugen, Filtern, Manipulieren und Kombinieren von *Tupelströmen*

- dabei evtl. Verwendung von Indexstrukturen (“Wegweiser”), Hashing und Sortier-Schritten
- wünschenswert: möglichst wenig auf Festplatte zwischenspeichern
- Operationen: Operationen der relationalen Algebra, Sortieren, Duplikatelimination, ...

Folie 216

Verwaltung des Sekundärspeichers



Hier (der Einfachheit halber): 1 Plattenzugriff $\hat{=}$ Lesen eines Blocks bzw. einer Seite

Folie 217

Wichtige Parameter einer Datenbankrelation I

- n_I : Anzahl der Tupel in Relation I
- s_I : (mittlere) Größe eines Tupels aus I
- f_I : Blockungsfaktor (“Wie viele Tupel aus I passen in einen Block?”)

$$f_I \approx \frac{\text{Blockgröße}}{s_I}$$

- b_I : Anzahl der Blöcke (Seiten) der Festplatte, die Tupel aus I beinhalten

$$b_I \approx \frac{n_I}{f_I}$$

Hier (der Einfachheit halber):

Lesen eines Blocks (bzw. einer Seite) $\hat{=}$ 1 Zugriff auf Platte

Folie 218

Operationen der relationalen Algebra

Selektion $\sigma_F(I)$:

- Selektion meist als Filter auf einem Tupelstrom:
 $\mathcal{O}(b_I)$ Zugriffe auf Festplatte; $\mathcal{O}(n_I)$ Schritte insgesamt
- evtl. Verwendung eines Index;
dann schneller, falls nur sehr wenige Tupel im Ergebnis

Folie 219

Projektion $\pi_{j_1, \dots, j_k}(I)$:

- 2 Komponenten:
 - Ändern der einzelnen Tupel (Auswahl und Reihenfolge von Spalten)
 - Duplikatelimination
- Tupeländerung: als Filter auf einem Tupelstrom
 $\mathcal{O}(b_I)$ Zugriffe auf Festplatte; $\mathcal{O}(n_I)$ Schritte insgesamt
Dabei können Duplikate “entstehen”
- Duplikatelimination:
 - in SQL i.d.R. nicht verlangt (außer bei `SELECT DISTINCT`)
 - sind die Tupel sortiert, so können Duplikate durch einen Scan leicht erkannt werden
 - Sortieren: durch Merge-Sort möglich mit $\mathcal{O}(b_I \cdot \log b_I)$ Plattenzugriffen und $\mathcal{O}(n_I \cdot \log n_I)$ Schritten insgesamt
 - Alternative: Hashing
 - * Abbilden der Tupel durch eine Hash-Funktion

- * \rightsquigarrow Duplikate werden auf denselben Wert abgebildet und dadurch erkannt
- * bei idealer Hash-Funktion: lineare Zeit

Folie 220

Binäre Operationen auf zwei Relationen I und J : \cup , $-$, \times , \bowtie_{θ} , \ltimes_{θ}

- *Nested-Loops-Methode*: (Schleifeniteration)
für jedes Tupel $t \in I$ (bzw. jede Seite) wird die gesamte Relation J durchlaufen
 $\mathcal{O}(b_I \cdot b_J)$ Plattenzugriffe; $\mathcal{O}(n_I \cdot n_J)$ Schritte insgesamt
- *Merge-Methode*: (weniger sinnvoll für \times)
 I und J sortiert \rightsquigarrow schrittweise in der vorgegebenen Tupelreihenfolge durchlaufen;
Für \ltimes_{θ} : $\mathcal{O}(b_I + b_J)$ Plattenzugriffe; $\mathcal{O}(n_I + n_J)$ Gesamtschritte
Evtl. vorher nötig: Sortieren von I und/oder J (durch Merge-Sort)
 $\mathcal{O}(b_I \cdot \log b_I)$ und/oder $\mathcal{O}(b_J \cdot \log b_J)$ Plattenzugriffe;
 $\mathcal{O}(n_I \cdot \log n_I)$ und/oder $\mathcal{O}(n_J \cdot \log n_J)$ Gesamtschritte
- *Hash-Methode*: (weniger sinnvoll für \times)
die kleinere der beiden Relationen in Hash-Tabelle;
Tupel der zweiten Relation finden ihren Vergleichspartner mittels Hash-Funktion;
bei idealer Hash-Funktion: Aufwand $\mathcal{O}(n_I + n_J)$

Folie 221

Beispiel für Merge-Technik

Berechne $I \bowtie_{\theta} J$ für Join-Bedingung $\theta := x_1=y_4 \wedge x_2=y_1$

1. Sortiere I lexikographisch nach "1-te Spalte; 2-te Spalte"
2. Sortiere J lexikographisch nach "4-te Spalte; 1-te Spalte"
3. Seien t und s die ersten Tupel von I und J
4. Falls $(t_1, t_2) < (s_4, s_1)$, so lies nächstes Tupel t aus I .
5. Falls $(t_1, t_2) > (s_4, s_1)$, so lies nächstes Tupel s aus J .

6. Falls $(t_1, t_2) = (s_4, s_1)$, so gib die Tupel (t, s) und (t, s') für alle Nachfolger s' von s in J mit $(s'_4, s'_1) = (s_4, s_1)$ aus
7. Lies nächstes Tupel t aus I und gehe zu Zeile 4.

Aufwand für Zeilen 3–7:

- falls alle Tupel den gleichen Wert in den Spalten 1,2 bzw. 4,1 haben:
ca. $n_I \cdot n_J$ Gesamtschritte
- falls alle Tupel aus J in (s_4, s_1) unterschiedliche Werte haben:
ca. $n_I + n_J$ Gesamtschritte :-)
- bei Semijoin \bowtie_θ statt Theta-Join \Join_θ reichen immer $n_I + n_J$ Gesamtschritte

Folie 222

Anfrageauswertung

Proposition 6.4. *Das Auswertungsproblem für die relationale Algebra lässt sich in Zeit $(k+n)^{\mathcal{O}(k)}$ lösen.*

Beweis:

Zeige per Induktion nach dem Aufbau von Anfragen der relationalen Algebra, dass für jede Anfrage Q der Länge k und jede Datenbank \mathbf{I} der Größe n gilt:

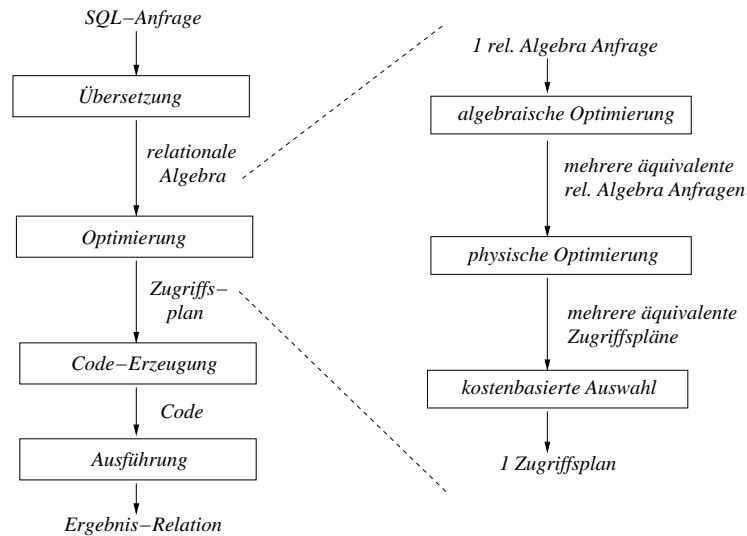
$$(1) \quad \|\llbracket Q \rrbracket(\mathbf{I})\| \leq (k+n)^k$$

(2) Q kann auf \mathbf{I} in $\mathcal{O}((k+n)^{2k})$ Elementarschritten ausgewertet werden.

Details: *Übung.*

Folie 223

Anfragebearbeitung durch ein DBMS



Folie 224

Ziel der Optimierung

- möglichst schnelle Auswertung der Anfrage
- möglichst wenige Zugriffe auf Festplatte

Grundregeln:

- (1) Selektionen so früh wie möglich
- (2) auch Projektionen früh, aber evtl. Duplikatelimination vermeiden
- (3) Basisoperationen zusammenfassen und wenn möglich ohne Zwischenspeicherung realisieren
(Bsp: \bowtie_{θ} besser als \times ; \times_{θ} besser als \bowtie_{θ})
- (4) Redundante Operationen oder leere Zwischenrelationen entfernen
- (5) Zusammenfassung gleicher Teilausdrücke:
Wiederverwendung von Zwischenergebnissen

Folie 225

Anfrageauswertung an einem Beispiel

Anfrage:

Welche Kinos (Name + Adresse) spielen einen Film von "James Cameron"?

In SQL:

```
SELECT Kinos.Name, Kinos.Adresse
FROM Kinos, Filme, Programm
WHERE Kinos.Name = Programm.Kino and
      Programm.Titel = Filme.Titel and
      Filme.Regie = 'James Cameron'
```

Direkte Übersetzung in relationale Algebra:

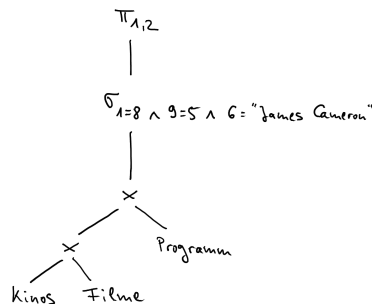
$$\pi_{1,2} \left(\sigma_{\substack{1=8 \wedge 9=5 \wedge \\ 6="James\ Cameron"}} (Kinos \times Filme \times Programm) \right)$$

Original-Anfrage

Folie 226

$$\pi_{1,2} \left(\sigma_{\substack{1=8 \wedge 9=5 \wedge \\ 6="James\ Cameron"}} (Kinos \times Filme \times Programm) \right)$$

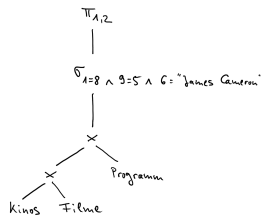
dargestellt als Anfrage-Baum:



Folie 227

Anfrage auswerten auf folgender Beispiel-Datenbank:

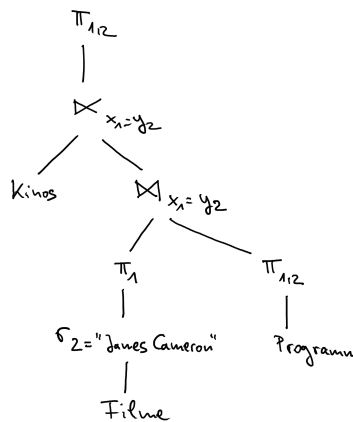
- *Filme*: 10.000 Tupel auf 200 Seiten (je 50 pro Seite); je 5 Tupel pro Film, 10 Filme von James Cameron
- *Programm*: 200 Tupel auf 4 Seiten (je 50 pro Seite); davon 3 Cameron-Filme in 4 Kinos
- *Kinos*: 100 Tupel auf 2 Seiten (je 50 pro Seite)



Anfrage-Baum: Direkte Auswertung dieser Anfrage führt zu über 10.000.000 Plattenzugriffen.

Folie 228

Viel besserer Plan:



Auswertung dieses Plans in unserer Beispiel-Datenbank führt zu weniger als 250 Plattenzugriffen.

Folie 229

Heuristische Optimierung

- Die heuristische Optimierung wendet allgemeine Regeln zur Umformung einer Anfrage der relationalen Algebra in eine äquivalente Anfrage an, die zu einem vermutlich effizienteren Auswertungsplan führt
- Grundregel: Selektionen so früh wie möglich
- Projektionen auch früh, aber evtl. Duplikatelimination vermeiden
- Anwendung von algebraischen Umformungsregeln
- Ziel: Operationen verschieben, um kleinere Zwischenergebnisse zu erhalten; wenn möglich Redundanzen erkennen

Einige algebraische Umformungsregeln:

- (1) Kartesische Produkte und Joins sind kommutativ und assoziativ:

$$\begin{aligned} Q_1 \bowtie_{\theta} Q_2 &\longleftrightarrow Q_2 \bowtie_{\tilde{\theta}} Q_1 \\ (Q_1 \bowtie_{\theta_1} Q_2) \bowtie_{\theta_2} Q_3 &\longleftrightarrow Q_1 \bowtie_{\tilde{\theta}_1} (Q_2 \bowtie_{\tilde{\theta}_2} Q_3) \end{aligned}$$

($\tilde{\theta}$ entsteht aus θ durch "Zurückrechnen" der Spaltennummern)

- (2) Ketten von Selektionen (bzw. Projektionen) zusammenfassen:

$$\begin{aligned} \sigma_{F_1}(\sigma_{F_2}(Q)) &\longleftrightarrow \sigma_{F_1 \wedge F_2}(Q) \longleftrightarrow \sigma_{F_2}(\sigma_{F_1}(Q)) \\ \pi_X(\pi_Y(Q)) &\longleftrightarrow \pi_{\tilde{X}}(Q) \end{aligned}$$

(\tilde{X} entsteht aus X durch "Zurückrechnen" der Spaltennummern)

- (3) Vertauschen von Selektion und Join:

$$\sigma_{F_1 \wedge F_2 \wedge F_3}(Q_1 \times Q_2) \longleftrightarrow \sigma_{F_1}(Q_1) \bowtie_{\theta_3} \sigma_{\tilde{F}_2}(Q_2),$$

wobei die Selektionsbed. F_1 (F_2) sich nur auf Spalten von Q_1 (Q_2) bezieht und F_3 Spalten von Q_1 mit Spalten von Q_2 vergleicht. \tilde{F}_2 und θ_3 entstehen aus F_2 und F_3 durch "Zurückrechnen" der Spaltennummern.

- (4) Einführung von Semijoins, falls X nur Spalten von Q_1 beinhaltet:

$$\pi_X(Q_1 \bowtie_{\theta} Q_2) \longleftrightarrow \pi_X(Q_1 \ltimes_{\theta} Q_2)$$

- (5) Vertauschen von Selektion und Vereinigung bzw. Differenz:

$$\begin{aligned} \sigma_F(Q_1 \cup Q_2) &\longleftrightarrow \sigma_F(Q_1) \cup \sigma_F(Q_2) \\ \sigma_F(Q_1 - Q_2) &\longleftrightarrow \sigma_F(Q_1) - \sigma_F(Q_2) \end{aligned}$$

- (6) Analog: Vertauschen von Projektion und Vereinigung bzw. Differenz.

- (7) Vertauschen von Projektion und Selektion unter bestimmten Bedingungen

- (8) Vertauschen von Projektion und Join unter bestimmten Bedingungen

- (9) Löschen von Redundanzen:

$$Q \cup Q \longrightarrow Q, \quad Q \cap Q \longrightarrow Q, \quad Q \bowtie Q \longrightarrow Q$$

- (10) Löschen leerer Zwischenergebnisse:

$$\begin{aligned} Q - Q &\longrightarrow \emptyset, & Q \cap \emptyset &\longrightarrow \emptyset, & Q \cup \emptyset &\longrightarrow Q, \\ Q \bowtie \emptyset &\longrightarrow \emptyset, \\ Q - \emptyset &\longrightarrow Q, & \emptyset - Q &\longrightarrow \emptyset \end{aligned}$$

- (11) ... usw. ...

Wunschliste für bessere Optimierung:

- zum “Löschen leerer Zwischenergebnisse”:
Test, ob eine gegebene (Teil-)Anfrage Q nicht erfüllbar ist ($“Q \equiv \emptyset”$)
- zum “Löschen von Redundanzen”:
Test, ob zwei (Teil-)Anfragen Q und P äquivalent sind ($“Q \equiv P”$)

Wir kennen bereits:

- Algorithmen zum Lösen dieser Probleme für konjunktive Anfragen

Im nächsten Kapitel:

- Nicht-Entscheidbarkeit dieser Probleme für allgemeine Anfragen der relationalen Algebra

Folie 233

Vorgehensweise eines Optimierers in einem DBMS

- Erzeugung verschiedener rel. Algebra Ausdrücke:
unter Verwendung heuristischer Optimierungsregeln
- Erzeugung von verschiedenen Auswertungsplänen:
Anfrage-Bäume, erweitert um Informationen über zu verwendende Zugriffsstrukturen und Algorithmen für die einzelnen Operationen
- Abschätzung der Kosten für jeden erzeugten Auswertungsplan:
unter Verwendung von statistischen Informationen über die Daten
(\leadsto kostenbasierte Optimierung)
- Auswahl des am günstigsten erscheinenden Plans

Generell:

Je häufiger die Anfrage ausgewertet werden soll, desto mehr Aufwand sollte für die Optimierung verwendet werden.

Folie 234

Join-Reihenfolge

- Joins sind kommutativ und assoziativ
 \rightsquigarrow Änderung der Klammerung bzw. Reihenfolge von
 Join-Operationen ändert nicht das Ergebnis der Anfrage (modulo
 Ändern der Spalten-Reihenfolge)
- Aber: Die Größe der Zwischenergebnisse und somit der
 Auswertungs-Aufwand kann sich drastisch ändern.
- Unter Umständen lässt sich sogar die Anzahl der Joins verkleinern
 (das haben wir bereits bei der Minimierung von konjunktiven
 Anfragen gesehen)
- Klassische Vorgehensweise in DBMS: Nur Auswertungspläne
 betrachten, die Joins von links nach rechts klammern
 (“left-deep-trees”)
 \rightsquigarrow durch Umordnen sind immerhin alle Reihenfolgen möglich (immer
 noch exponentiell viele Möglichkeiten)
- (Es ist bekannt, dass diese Einschränkung nicht immer sinnvoll und
 nötig ist)

Jetzt: Heuristische Join-Optimierung bzgl. left-deep-trees

Folie 235

Beispiele

Beispiel 6.5. R : 2-stellige Relation mit Attributen A, B und 1.000 Tupeln

S : 2-stellige Relation mit Attributen C, D und 10.000 Tupeln

T : 2-stellige Relation mit Attributen B, C und 100 Tupeln

Pro Tupel $(b, c) \in T$ gibt es ein Tupel $(\cdot, b) \in R$ und ein Tupel $(c, \cdot) \in S$.

Anfrage: $Ans(x) \leftarrow R(x, x_1), S(x_2, x_3), T(x_1, x_2)$

Aufwand bei Links-nach-rechts-Auswertung:

10.000.000 Tupel nach erstem Join, 100 Tupel nach zweitem Join

Andere Join-Reihenfolge: $Ans(x) \leftarrow R(x, x_1), T(x_1, x_2), S(x_2, x_3)$

Aufwand bei Links-nach-rechts-Auswertung:

100 Tupel nach erstem Join, 100 Tupel nach zweitem Join

Folie 236

Beispiel 6.6. R : 2-stellige Relation mit Attributen A, B und 1.000 Tupeln

S : 2-stellige Relation mit Attributen C, D und 10.000 Tupeln

T : 2-stellige Relation mit Attributen B, C und 100 Tupeln

Pro Tupel $(b, c) \in T$ gibt es ein Tupel $(\cdot, b) \in R$ und ein Tupel $(c, \cdot) \in S$.

Für die Konstante d gibt es nur 1 Tupel (\cdot, d) in S .

Anfrage: $Ans(x) \leftarrow R(x, x_1), T(x_1, x_2), S(x_2, d)$

Aufwand bei Links-nach-rechts-Auswertung:

100 Tupel nach erstem Join, max. 1 Tupel nach zweitem Join

Andere Join-Reihenfolge: $Ans(x) \leftarrow S(x_2, d), T(x_1, x_2), R(x, x_1)$

Aufwand bei Links-nach-rechts-Auswertung:

max. 1 Tupel nach erstem Join, max. 1 Tupel nach zweitem Join

Folie 237

Noch ein Beispiel: a und b seien jetzt Konstanten, d.h. $a, b \in \text{dom}$

- Auswertung von links nach rechts
- Anfrage: $Ans(z) \leftarrow R(v, w, y), v \leq x, S(x, y, z), P(a, v), Q(b, w, x)$
- Besserer Auswertungsplan:
 $Ans(z) \leftarrow P(a, v), Q(b, w, x), R(v, w, y), S(x, y, z), v \leq x,$
- Denn:
 - wahrscheinlich wenige Tupel der Form (a, \cdot) in P
 - wahrscheinlich wenige Tupel der Form (b, \cdot, \cdot) in Q
 - wahrscheinlich wenige Tupel, die $P(a, v), Q(b, w, x), R(v, w, y)$ erfüllen

Heuristik (“Sideways-Information-Passing”, kurz: SIP):

- Relations-Atome mit Konstanten zuerst auswerten
- Wenn möglich, Relations-Atome erst dann, wenn weiter links schon eine ihrer Variablen steht.
- Vergleichsoperatoren (\leq , $<$) möglichst erst dann verwenden, wenn beide Variablen schon verwendet wurden.

Folie 238

SIP-Graph und SIP-Strategie

Definitionen:

- Regel $Ans(u) \leftarrow R_1(u_1), \dots, R_k(u_k), E_1, \dots, E_\ell, C_1, \dots, C_m$
wobei E_i Vergleich mit =
und C_i Vergleich mit $<$ oder \leq (dafür sei $<$ eine lineare Ordnung auf **dom**).
- *SIP-Graph*
 - Knotenmenge: Rel.-Atome $R_1(u_1), \dots, R_k(u_k)$ und “=”-Atome E_1, \dots, E_ℓ
 - Kante zwischen zwei Knoten, falls diese (mind.) eine Variable gemeinsam haben
 - Knoten ist *markiert*, falls er mind. eine Konstante enthält
- Falls der SIP-Graph zusammenhängend ist, so ist eine *SIP-Strategie* eine Anordnung $A_1, \dots, A_{k+\ell+m}$ der Atome, so dass für jedes $j > 1$ gilt:
 - A_j ist ein *markierter* Knoten des SIP-Graphen, oder
 - A_j ist ein Knoten des SIP-Graphen und es gibt ein $j' < j$, so dass es im SIP-Graph eine Kante zwischen A_j und $A_{j'}$ gibt, oder
 - A_j ist ein C_i und für jede Variable x in C_i gibt es $j' < j$, so dass $A_{j'}$ ein Knoten des SIP-Graphen ist, in dem x vorkommt
- Außerdem: Falls ex. $R_i(u_i)$ od. E_i mit mind. einer Konstanten, so ist A_1 ein solches Atom

- Falls der SIP-Graph nicht zusammenhängend ist: SIP-Strategie für jede Zusammenhangskomponente.

Folie 239

Beispiele: a und b seien Konstanten, d.h. $a, b \in \mathbf{dom}$.

- $Ans(z) \leftarrow P(a, v), Q(b, w, x), R(v, w, y), S(x, y, z), v \leq x$
ist in der Reihenfolge einer SIP-Strategie
- $Ans(z) \leftarrow R(v, w, y), v \leq x, S(x, y, z), P(a, v), Q(b, w, x)$
ist nicht in der Reihenfolge einer SIP-Strategie

Bemerkungen:

- Zu jeder regelbasierten konjunktiven Anfrage (evtl. mit $=$ oder \leq ; dann aber bereichsbeschränkt) gibt es eine SIP-Strategie.
- Eine SIP-Strategie für eine gegebene Anfrage lässt sich in Zeit polynomiell in der Größe der Anfrage erzeugen.
- Wird eine Variable weiter rechts nicht mehr benötigt, so kann sie beim Auswerten der Anfrage aus dem Zwischenergebnis “heraus projiziert” werden.

Kapitel 7

Relationenkalkül

7.1 CALC_{nat} , $\text{CALC}_{\text{adom}}$ und CALC_{di}

Folie 240

Einleitung

Algebraisch	... äquivalent dazu ...	Logik-basiert
SPC-Algebra		konjunktiver Kalkül
Boolesche Semijoin-Anfragen		GF(CQ)-Sätze
relationale Algebra		Relationenkalkül

Der *konjunktive Kalkül* basiert auf einem *Fragment der Logik erster Stufe*.

Der *Relationenkalkül* basiert auf der vollen *Logik erster Stufe* (kurz: FO).

(“FO” steht für “*first-order logic*”)

Folie 241

Die Logik erster Stufe — $\text{FO}[\mathbf{S}]$

Definition 7.1. Sei \mathbf{S} ein Datenbankschema.

Die Menge $\text{FO}[\mathbf{S}]$ aller Formeln der *Logik erster Stufe* über \mathbf{S} ist induktiv wie folgt definiert:

- (A) “*Relations-Atome*”: $R(v_1, \dots, v_r)$ gehört zu $\text{FO}[\mathbf{S}]$,
für alle $R \in \mathbf{S}$, $r := \text{ar}(R)$ und $v_1, \dots, v_r \in \text{var} \cup \mathbf{dom}$.
- (G) “*Gleichheits-Atome*”: $v_1 = v_2$ gehört zu $\text{FO}[\mathbf{S}]$, für alle
 $v_1, v_2 \in \text{var} \cup \mathbf{dom}$.

(BK) “*Boolesche Kombinationen*”: Falls $\varphi_1 \in \text{FO}[\mathbf{S}]$ und $\varphi_2 \in \text{FO}[\mathbf{S}]$, so gehört auch jede der folgenden fünf Formeln zu $\text{FO}[\mathbf{S}]$:

$\neg\varphi_1$, $(\varphi_1 \wedge \varphi_2)$, $(\varphi_1 \vee \varphi_2)$, $(\varphi_1 \rightarrow \varphi_2)$ und $(\varphi_1 \leftrightarrow \varphi_2)$.

(Q) “*Quantoren*”: Ist $\varphi_1 \in \text{FO}[\mathbf{S}]$ und ist $x \in \text{var}$, so gehören auch die beiden folgenden Formeln zu $\text{FO}[\mathbf{S}]$: $\exists x \varphi_1$ und $\forall x \varphi_1$.

Bemerkung: Benutzen wir die Notation aus der Vorlesung „*Logik in der Informatik*“, so ist $\text{FO}[\mathbf{S}]$ genau die Menge aller Formeln aus $\text{FO}[\sigma]$, für die Signatur $\sigma := \mathbf{S} \cup \mathbf{dom}$, wobei jedes Element aus \mathbf{dom} als Konstantensymbol aufgefasst wird, das stets „mit sich selbst“ interpretiert wird).

Folie 242

Notationen und Anmerkungen

- Manchmal werden wir Formeln der Form

– $(\varphi_1 \vee \varphi_2)$ als Abkürzung für $\neg(\neg\varphi_1 \wedge \neg\varphi_2)$,

– $(\varphi_1 \rightarrow \varphi_2)$ als Abkürzung für $(\neg\varphi_1 \vee \varphi_2)$

– $(\varphi_1 \leftrightarrow \varphi_2)$ als Abkürzung für $(\varphi_1 \wedge \varphi_2) \vee (\neg\varphi_1 \wedge \neg\varphi_2)$

– $\forall x \varphi_1$ als Abkürzung für $\neg \exists x \neg \varphi_1$

auffassen.

- $\text{adom}(\varphi)$ bezeichnet die Menge aller Konstanten (also Elemente aus \mathbf{dom}), die in φ vorkommen.

$\text{Var}(\varphi)$ bezeichnet die Menge aller Variablen (also Elemente aus var , die in φ vorkommen).

$\text{frei}(\varphi)$ bezeichnet die Menge aller Variablen, die *frei* in φ vorkommen. D.h.

– $\text{frei}(R(v_1, \dots, v_r)) := \{v_1, \dots, v_r\} \cap \text{var}$

– $\text{frei}(v_1=v_2) := \{v_1, v_2\} \cap \text{var}$

– $\text{frei}(\neg\varphi_1) := \text{frei}(\varphi_1)$

– $\text{frei}((\varphi_1 * \varphi_2)) := \text{frei}(\varphi_1) \cup \text{frei}(\varphi_2)$ (für alle $* \in \{\wedge, \vee, \rightarrow, \leftrightarrow\}$)

– $\text{frei}(\exists x \varphi_1) := \text{frei}(\forall x \varphi_1) := \text{frei}(\varphi_1) \setminus \{x\}$

- Eine *Belegung für φ* ist eine Abbildung $\beta : \text{frei}(\varphi) \rightarrow \mathbf{dom}$.
- Sei $\mathbf{d} \subseteq \mathbf{dom}$. Eine *Belegung für φ in \mathbf{d}* ist eine Abbildung $\beta : \text{frei}(\varphi) \rightarrow \mathbf{d}$.

Folie 243

Natürliche Semantik von FO[S] und CALC

Wir werden verschiedene Varianten der Semantik von FO[S] betrachten. Hier zunächst die *natürliche Semantik*:

Definition 7.2 (natürliche Semantik von FO[S]).

- *Zur Erinnerung:* Einer Datenbank \mathbf{I} vom Schema \mathbf{S} ordnen wir die logische Struktur $\mathcal{A}_{\mathbf{I}} := (\mathbf{dom}, (\mathbf{I}(R))_{R \in \mathbf{S}}, (c)_{c \in \mathbf{dom}})$ zu.
- Ist \mathbf{I} eine Datenbank vom Schema \mathbf{S} und β eine Belegung für φ (also eine Abbildung $\beta : \text{frei}(\varphi) \rightarrow \mathbf{dom}$), so sagen wir “ \mathbf{I} erfüllt φ unter β ” und schreiben $\mathbf{I} \models \varphi[\beta]$ bzw. $(\mathbf{I}, \beta) \models \varphi$, um auszudrücken, dass $(\mathcal{A}_{\mathbf{I}}, \beta) \models \varphi$.

Definition 7.3 (Relationenkalkül CALC). Sei \mathbf{S} ein Datenbankschema.

Eine Anfrage des *Relationenkalküls* CALC[S] ist von der Form

$$\{ (e_1, \dots, e_r) : \varphi \},$$

wobei $\varphi \in \text{FO}[\mathbf{S}]$, $r \geq 0$ und (e_1, \dots, e_r) ein freies Tupel ist (d.h.

$e_1, \dots, e_r \in \text{var} \cup \mathbf{dom}$), so dass $\text{frei}(\varphi) = \{e_1, \dots, e_r\} \cap \text{var}$.

Wir setzen $\text{adom}(Q) := \text{adom}(\varphi) \cup (\{e_1, \dots, e_r\} \cap \mathbf{dom})$.

Semantik: Einer CALC[S]-Anfrage Q der obigen Form ordnen wir die folgende “Anfragefunktion” $\llbracket Q \rrbracket_{\text{nat}}$ zu (f.a. $\mathbf{I} \in \text{inst}(\mathbf{S})$):

$$\llbracket Q \rrbracket_{\text{nat}}(\mathbf{I}) := \{ \beta((e_1, \dots, e_r)) : \beta \text{ ist eine Belegung für } \varphi \text{ mit } \mathbf{I} \models \varphi[\beta] \}$$

Folie 244

Beispiele für CALC-Anfragen

- In welchen Filmen hat “George Clooney” mitgespielt, aber nicht selbst Regie geführt?

$$\left\{ (x_{\text{Titel}}) : \exists x_{\text{Regie}} \left(\text{Filme}(x_{\text{Titel}}, x_{\text{Regie}}, \text{“George Clooney”}) \wedge \neg \text{Filme}(x_{\text{Titel}}, \text{“George Clooney”}, \text{“George Clooney”}) \right) \right\}$$

- Welche Filme laufen nur zu 1 Uhrzeit?

$$\left\{ (x_{\text{Titel}}) : \exists x_{\text{Kino}} \exists x_{\text{Zeit}} \left(\text{Programm}(x_{\text{Kino}}, x_{\text{Titel}}, x_{\text{Zeit}}) \wedge \forall y_{\text{Kino}} \forall y_{\text{Zeit}} (\text{Programm}(y_{\text{Kino}}, x_{\text{Titel}}, y_{\text{Zeit}}) \rightarrow y_{\text{Zeit}} = x_{\text{Zeit}}) \right) \right\}$$

- Welche Filme haben nur Schauspieler, die schon mal in einem Film von “Stephen Spielberg” mitgespielt haben?

$$\left\{ (x_T) : \exists x_R \exists x_S \left(\text{Filme}(x_T, x_R, x_S) \wedge \forall y_S (\text{Filme}(x_T, x_R, y_S) \rightarrow \exists y_T \text{Filme}(y_T, \text{“Stephen Spielberg”}, y_S)) \right) \right\}$$

Frage: Was ist mit ... ?

$$\left\{ (x_T) : \forall y_S (\exists x_R \text{Filme}(x_T, x_R, y_S) \rightarrow \exists y_T \text{Filme}(y_T, \text{“Stephen Spielberg”}, y_S)) \right\}$$

Folie 245

Unsichere CALC-Anfragen

Beispiele:

$$(1) \quad Q_1 := \{ (x_{\text{Schausp}}) : \neg \text{Filme}(\text{“Alien”}, \text{“Ridley Scott”}, x_{\text{Schausp}}) \}$$

$$(2) \quad Q_2 := \{ (x_S, y_T) : \text{Filme}(\text{“Alien”}, \text{“Ridley Scott”}, x_S) \vee \text{Filme}(y_T, \text{“George Clooney”}, \text{“George Clooney”}) \}$$

$$(3) \quad Q_3 :=$$

$$\left\{ (x_T) : \forall y_S (\exists x_R \text{Filme}(x_T, x_R, y_S) \rightarrow \exists y_T \text{Filme}(y_T, \text{“Stephen Spielberg”}, y_S)) \right\}$$

Auswertung mit $[\cdot]_{\text{nat}}$: Q_1 liefert alle (unendlich vielen) $a \in \mathbf{dom}$, die nicht als Schauspieler im Film “Alien” mitgespielt haben.

Q_2 liefert alle (unendlich vielen) Tupel der Form (a, b) , für die a ein Schauspieler in “Alien” und b ein beliebiges Element in \mathbf{dom} ist oder b ein Film von und mit “George Clooney” und a ein beliebiges Element aus \mathbf{dom} .

Q_3 liefert alle Filme, die nur solche Schauspieler haben, die schon mal mit “Stephen Spielberg” gearbeitet haben, aber auch **alle Elemente aus dom, die nicht Titel eines Films sind**.

Problem: Gemäß der bisherigen Definition der Semantik liefern die Anfragen Q_1 , Q_2 und Q_3 stets “unendliche Ergebnisse” die gemäß unserer Definition *keine Datenbank-Relationen* sind (da die stets endlich sein müssen). Solche Anfrage heißen *unsicher*.

Folie 246

Eine weitere problematische Anfrage

$$Q_4 := \{ (x) : \forall y R(x, y) \}$$

Q_4 liefert als Ergebnis stets die leere Relation, d.h.
 $\llbracket Q_4 \rrbracket_{nat}(\mathbf{I}) = \emptyset$, für alle $\mathbf{I} \in inst(\mathbf{S})$.

Dies ist unschön.

Daher (und weil manche Anfragen unsicher sind in der nat. Semantik), betrachten wir im Folgenden andere Varianten der Semantik.

Folie 247

Relativierte Semantik von FO[S] und CALC[S]

Definition 7.4. Sei \mathbf{S} ein Datenbankschema.

- (a) Eine *relativierte Datenbank* über \mathbf{S} ist ein Paar (\mathbf{d}, \mathbf{I}) , wobei \mathbf{I} eine Datenbank vom Schema \mathbf{S} ist und \mathbf{d} eine Menge, so dass $\text{adom}(\mathbf{I}) \subseteq \mathbf{d} \subseteq \text{dom}$.
- (b) Einer relativierten Datenbank (\mathbf{d}, \mathbf{I}) über \mathbf{S} ordnen wir die logische Struktur $\mathcal{A}_{(\mathbf{d}, \mathbf{I})} := (\mathbf{d}, (\mathbf{I}(R))_{R \in \mathbf{S}}, (c)_{c \in \mathbf{d}})$ zu.
- (c) Eine FO[S]-Formel φ heißt *interpretierbar über (\mathbf{d}, \mathbf{I})* , falls $\text{adom}(\varphi) \subseteq \mathbf{d}$.
 Eine CALC[S]-Anfrage Q heißt *interpretierbar über (\mathbf{d}, \mathbf{I})* , falls $\text{adom}(Q) \subseteq \mathbf{d}$.
- (d) Ist (\mathbf{d}, \mathbf{I}) eine relativierte Datenbank über \mathbf{S} , ist φ eine FO[S]-Formel, die interpretierbar ist über (\mathbf{d}, \mathbf{I}) , und ist β eine Belegung für φ in \mathbf{d} (also eine Abbildung $\beta : \text{frei}(\varphi) \rightarrow \mathbf{d}$), so sagen wir “ \mathbf{I} erfüllt φ unter β relativ zu \mathbf{d} ” und schreiben $\mathbf{I} \models_{\mathbf{d}} \varphi[\beta]$ bzw. $(\mathbf{I}, \beta) \models_{\mathbf{d}} \varphi$, um auszudrücken, dass $(\mathcal{A}_{(\mathbf{d}, \mathbf{I})}, \beta) \models \varphi$.
- (e) Ist Q eine CALC[S]-Anfrage der Form $\{(e_1, \dots, e_r) : \varphi\}$ und (\mathbf{d}, \mathbf{I}) eine relativierte Datenbank über \mathbf{S} , über der Q interpretierbar ist, so ist

$$\llbracket Q \rrbracket_{\mathbf{d}}(\mathbf{I}) := \{ \beta((e_1, \dots, e_r)) : \beta \text{ ist eine Belegung für } \varphi \text{ über } \mathbf{d} \text{ mit } \mathbf{I} \models_{\mathbf{d}} \varphi[\beta] \}$$

(“Relativiert” soll andeuten, dass Quantifizierung relativiert wird auf Elemente aus \mathbf{d} .)

Folie 248

Ein Beispiel

Beispiel 7.5.

- \mathbf{S} enthalte ein 1-stelliges Relationssymbol R
- \mathbf{I} sei die Datenbank mit
 $\mathbf{I}(R) = \{(a), (b), (c)\}$ und $\mathbf{I}(S) = \emptyset$ für alle $S \in \mathbf{S} \setminus \{R\}$.
- Q sei die CALC[\mathbf{S}]-Anfrage

$$Q := \{ (x) : (R(x) \wedge \exists y (\neg R(y) \wedge \forall z (R(z) \vee z=y))) \}$$

- Dann gilt:
 - $\llbracket Q \rrbracket_{\{a,b,c\}}(\mathbf{I}) = \emptyset$
 - $\llbracket Q \rrbracket_{\{a,b,c,d\}}(\mathbf{I}) = \{(a), (b), (c)\}$
 - $\llbracket Q \rrbracket_{\{a,b,c,d,e\}}(\mathbf{I}) = \emptyset$

Folie 249

Active Domain Semantik von FO[\mathbf{S}] und CALC[\mathbf{S}]

Definition 7.6.

- (a) Ist $\mathbf{I} \in \text{inst}(\mathbf{S})$ und $\varphi \in \text{FO}[\mathbf{S}]$, so ist $\text{adom}(\varphi, \mathbf{I}) := \text{adom}(\varphi) \cup \text{adom}(\mathbf{I})$.
 Ist $\mathbf{I} \in \text{inst}(\mathbf{S})$ und $Q \in \text{CALC}[\mathbf{S}]$, so ist
 $\text{adom}(Q, \mathbf{I}) := \text{adom}(Q) \cup \text{adom}(\mathbf{I})$.
- (b) Ist $\varphi \in \text{FO}[\mathbf{S}]$, \mathbf{I} eine Datenbank vom Schema \mathbf{S} und β eine Belegung für φ in $\text{adom}(\varphi, \mathbf{I})$ (also eine Abbildung $\beta : \text{frei}(\varphi) \rightarrow \text{adom}(\varphi, \mathbf{I})$), so sagen wir “ \mathbf{I} erfüllt φ unter β in der Active Domain Semantik” und schreiben $\mathbf{I} \models_{\text{adom}} \varphi[\beta]$ bzw. $(\mathbf{I}, \beta) \models_{\text{adom}} \varphi$, um auszudrücken, dass $\mathbf{I} \models_{\mathbf{d}} \varphi[\beta]$ für $\mathbf{d} := \text{adom}(\varphi, \mathbf{I})$ gilt.
- (c) Ist Q eine CALC[\mathbf{S}]-Anfrage der Form $\{(e_1, \dots, e_r) : \varphi\}$, so definiert Q in der Active Domain Semantik die folgende Anfragefunktion

$$\llbracket Q \rrbracket_{\text{adom}}(\mathbf{I}) := \llbracket Q \rrbracket_{\text{adom}(Q, \mathbf{I})}(\mathbf{I})$$

Folie 250

Bemerkungen zur natürlichen Semantik und zur Active Domain Semantik

- Die Semantik $\llbracket \cdot \rrbracket_{nat}$ ist vom Standpunkt der Logik aus “natürlich”; vom Standpunkt des Datenbanknutzers aber eher unnatürlich, da manche Anfragen unendliche Ergebnisse liefern (d.h. unsicher sind).
- Die Active Domain Semantik $\llbracket \cdot \rrbracket_{adom}$ liefert immer ein endliches Ergebnis; aber das Ergebnis kann sich (vom Standpunkt des Datenbanknutzers) auf unnatürliche Weise ändern, wenn ein neuer Wert in die Datenbank eingetragen wird (selbst wenn dieser scheinbar nichts mit der Anfrage zu tun hat).

Beispiel: $Q := \{ (x) : (R(x) \wedge \exists y(\neg R(y) \wedge \forall z(R(z) \vee z=y))) \}$

- Radikaler Schritt: Betrachte nur Anfragen, die *bereichsunabhängig* sind, d.h. bei denen es egal ist, ob sie in $\llbracket \cdot \rrbracket_{nat}$, $\llbracket \cdot \rrbracket_{adom}$ oder in $\llbracket \cdot \rrbracket_{\mathbf{d}}$ für irgendeine Menge \mathbf{d} mit $\text{adom}(Q, \mathbf{I}) \subseteq \mathbf{d} \subseteq \mathbf{dom}$ ausgewertet werden.

Definition 7.7 (bereichsunabhängige CALC[S]-Anfragen).

Eine Anfrage $Q \in \text{CALC}[\mathbf{S}]$ heißt *bereichsunabhängig* (domain independent), falls für jede Datenbank $\mathbf{I} \in \text{inst}(\mathbf{S})$ und alle $\mathbf{d}, \mathbf{d}' \subseteq \mathbf{dom}$ mit $\text{adom}(Q, \mathbf{I}) \subseteq \mathbf{d}, \mathbf{d}'$ gilt: $\llbracket Q \rrbracket_{\mathbf{d}}(\mathbf{I}) = \llbracket Q \rrbracket_{\mathbf{d}'}(\mathbf{I})$.

Folie 251

Beispiele 7.8. Beispiele für bereichsunabhängige Anfragen:

(a) Welche Filme laufen in mindestens 2 Kinos?

$$\left\{ (x_T) : \exists x_K \exists x_Z \exists y_K \exists y_Z (\text{Programm}(x_K, x_T, x_Z) \wedge \text{Programm}(y_K, x_T, y_Z) \wedge \neg x_K = y_K) \right\}$$

(b) In welchen Filmen hat “George Clooney” mitgespielt, aber nicht selbst Regie geführt?

$$\left\{ (x_{\text{Titel}}) : \exists x_{\text{Regie}} \left(\text{Filme}(x_{\text{Titel}}, x_{\text{Regie}}, \text{“George Clooney”}) \wedge \neg \text{Filme}(x_{\text{Titel}}, \text{“George Clooney”}, \text{“George Clooney”}) \right) \right\}$$

(c) Welche Filme laufen nur zu 1 Uhrzeit?

$$\left\{ (x_{\text{Titel}}) : \exists x_{\text{Kino}} \exists x_{\text{Zeit}} \left(\text{Programm}(x_{\text{Kino}}, x_{\text{Titel}}, x_{\text{Zeit}}) \wedge \forall y_{\text{Kino}} \forall y_{\text{Zeit}} (\text{Programm}(y_{\text{Kino}}, x_{\text{Titel}}, y_{\text{Zeit}}) \rightarrow y_{\text{Zeit}} = x_{\text{Zeit}}) \right) \right\}$$

Folie 252

- (d) Welche Filme haben nur Schauspieler, die schon mal in einem Film von “Stephen Spielberg” mitgespielt haben?

$$\left\{ (x_T) : \exists x_R \exists x_S \left(\text{Filme}(x_T, x_R, x_S) \wedge \forall y_S (\text{Filme}(x_T, x_R, y_S) \rightarrow \exists y_T \text{Filme}(y_T, \text{“Stephen Spielberg”}, y_S)) \right) \right\}$$

Nicht bereichunabhängig ist ...

$$\left\{ (x_{\text{Titel}}) : \forall y_S (\exists x_R \text{Filme}(x_{\text{Titel}}, x_R, y_S) \rightarrow \exists y_T \text{Filme}(y_T, \text{“Stephen Spielberg”}, y_S)) \right\}$$

Folie 253

CALC_{nat} , $\text{CALC}_{\text{adom}}$, CALC_{di}

Definition 7.9.

$\text{CALC}_{\text{nat}}[\mathbf{S}]$ bezeichnet die Klasse aller in der natürlichen Semantik $\llbracket \cdot \rrbracket_{\text{nat}}$ interpretierten $\text{CALC}[\mathbf{S}]$ -Anfragen.

$\text{CALC}_{\text{adom}}[\mathbf{S}]$ bezeichnet die Klasse aller in der Active Domain Semantik $\llbracket \cdot \rrbracket_{\text{adom}}$ interpretierten $\text{CALC}[\mathbf{S}]$ -Anfragen.

$\text{CALC}_{\text{di}}[\mathbf{S}]$ bezeichnet die Klasse aller bereichsunabhängigen $\text{CALC}[\mathbf{S}]$ -Anfragen (interpretiert in $\llbracket \cdot \rrbracket_{\text{adom}}$ oder, äquivalent dazu, in $\llbracket \cdot \rrbracket_{\text{nat}}$). (“di” steht für “domain independent”)

Satz 7.10 (). Die folgenden Anfragesprachen können genau dieselben Anfragefunktionen ausdrücken:

(a) CALC_{di}

(b) $\text{CALC}_{\text{adom}}$

(c) relationale Algebra (unbenannte oder benannte Perspektive).

Und für jedes feste Datenbankschema \mathbf{S} gilt: Jede Anfrage aus einer dieser Sprachen kann in polynomieller Zeit in äquivalente Anfragen der anderen Sprachen übersetzt werden.

Beweis: Siehe Tafel.

Folie 254

Bereichsunabhängigkeit — Pro & Contra CALC_{di}

Vorteile:

- logik-basierte Anfragesprache, die die “richtige” Ausdrucksstärke hat (äquivalent zur relationalen Algebra; insbesondere sind alle CALC_{di} -Anfragen “sicher”)
- keine unerwünschten Effekte, wie sie bei CALC_{adom} auftreten können (vgl. Beispiel 7.5).
- Ergebnis der Anfragen ist unabhängig davon, in welcher Semantik ($\llbracket \cdot \rrbracket_{adom}$ oder $\llbracket \cdot \rrbracket_{nat}$ oder $\llbracket \cdot \rrbracket_d$) sie interpretiert werden

Frage:

- Wie kann man bei einer gegebenen CALC-Anfrage Q feststellen, ob Q zu CALC_{di} gehört, d.h. ob Q bereichsunabhängig ist?

Antwort — großer Nachteil von CALC_{di} :

- Das ist *unentscheidbar*, d.h. es gibt keinen Algorithmus, der bei Eingabe einer beliebigen $\text{CALC}[\mathbf{S}]$ -Anfrage Q nach endlich vielen Schritten anhält und entscheidet, ob Q bereichsunabhängig ist oder nicht.
- Beweis: auf den nächsten Folien ...

Folie 255

Der Satz von Trakhtenbrot

Theorem 7.11 (Trakhtenbrot, 1950 (hier ohne Beweis)).

Sei \mathbf{S} ein Datenbankschema, das mindestens ein Relationssymbol enthält, dessen Stelligkeit ≥ 2 ist. Dann ist das folgende Problem unentscheidbar:

ERFÜLLBARKEITSPROBLEM FÜR $\text{FO}[\mathbf{S}]$ ÜBER DATENBANKEN

Eingabe: Ein $\text{FO}[\mathbf{S}]$ -Satz ψ

Frage: Gibt es eine DB $\mathbf{I} \in \text{inst}(\mathbf{S})$ mit $\mathbf{I} \models_{adom} \psi$?

Beweis: Siehe Vorlesung „Logik und Komplexität“.

Folie 256

Bereichsunabhängigkeit ist unentscheidbar

Satz 7.12. *Sei S ein Datenbankschema, das mindestens ein Relationssymbol R enthält, dessen Stelligkeit ≥ 2 ist. Dann ist das folgende Problem unentscheidbar:*

BEREICHSUNABHÄNGIGKEIT FÜR $\text{CALC}[S]$

Eingabe: Eine $\text{CALC}[S]$ -Anfrage Q

Frage: Ist Q bereichsunabhängig (d.h., gehört Q zu $\text{CALC}_{di}[S]$) ?

Beweis: Einfache Folgerung aus dem Satz von Trakhtenbrot; siehe Tafel.

7.2 Sicherer Relationenkalkül: CALC_{sr}

Folie 257

Motivation

- CALC_{di} ist eine logik-basierte Anfragesprache, die die *“richtige Ausdrucksstärke”* hat (nämlich dieselbe wie die relationale Algebra).
- Wegen der Unentscheidbarkeit (Satz 7.12) ist CALC_{di} als Anfragesprache für praktische Zwecke aber ungeeignet, da man bei Eingabe einer “Anfrage” aus $\text{CALC}[S]$ nicht feststellen kann, ob dies eine Anfrage in CALC_{di} ist.
- *Ziel jetzt:* Finde ein *syntaktisches Kriterium* für $\text{CALC}[S]$ -Anfragen, das
 - (a) algorithmisch leicht nachprüfbar ist,
 - (b) Bereichsunabhängigkeit garantiert und
 - (c) zu einer Anfragesprache führt, die immer noch dieselbe Ausdrucksstärke wie die relationale Algebra hat.

\leadsto *Sicherer Relationenkalkül* CALC_{sr} : entscheidbare Teilklasse von CALC_{di} , die dieselbe Ausdrucksstärke wie CALC_{di} hat.

Folie 258

Safe-Range Normalform (SRNF)

Für jede FO[S]-Formel φ sei $\text{SRNF}(\varphi)$ die FO[S]-Formel, die aus φ entsteht, indem man die folgenden Regeln so lange anwendet, bis keine Regel mehr anwendbar ist:

- Alle quantifizierten Variablen werden so umbenannt, dass sie paarweise verschieden sind und verschieden von den freien Variablen der Formel.
- Teilformeln der Form $\forall x\psi$ werden ersetzt durch $\neg\exists x\neg\psi$
- Implikationspfeile \rightarrow und “genau-dann-wenn”-Pfeile \leftrightarrow werden durch Kombinationen von \wedge, \vee, \neg ersetzt:
 - Teilformeln der Form $(\psi_1 \rightarrow \psi_2)$ werden ersetzt durch $(\neg\psi_1 \vee \psi_2)$
 - Teilformeln der Form $(\psi_1 \leftrightarrow \psi_2)$ werden ersetzt durch $((\psi_1 \wedge \psi_2) \vee (\neg\psi_1 \wedge \neg\psi_2))$
- Negationszeichen werden “nach innen geschoben”:
 - $\neg(\psi_1 \wedge \psi_2)$ wird ersetzt durch $(\neg\psi_1 \vee \neg\psi_2)$
 - $\neg(\psi_1 \vee \psi_2)$ wird ersetzt durch $(\neg\psi_1 \wedge \neg\psi_2)$
- “doppelte Negationszeichen” werden gelöscht:
 Teilformeln der Form $\neg\neg\psi$ werden ersetzt durch ψ

Offensichtlich gilt: Die Formel $\text{SRNF}(\varphi)$ ist äquivalent zur Formel φ .
 Falls $\varphi = \text{SRNF}(\varphi)$, so sagen wir: φ ist in *Safe-Range Normalform* (kurz: SRNf).

Folie 259

Beispiel für Transformation von φ zu $\text{SRNF}(\varphi)$

Beispiel 7.13.

$$\varphi := \exists x_R \exists x_S \left(\text{Filme}(x_T, x_R, x_S) \wedge \forall y_S (\text{Filme}(x_T, x_R, y_S) \rightarrow \exists y_T \text{Filme}(y_T, \text{“Stephen Spielberg”}, y_S)) \right)$$

$$\begin{aligned}
 & \rightsquigarrow \exists x_R \exists x_S \left(\text{Filme}(x_T, x_R, x_S) \wedge \right. \\
 & \quad \left. \neg \exists y_S \neg (\text{Filme}(x_T, x_R, y_S) \rightarrow \exists y_T \text{Filme}(y_T, \text{“Stephen Spielberg”}, y_S)) \right) \\
 & \rightsquigarrow \exists x_R \exists x_S \left(\text{Filme}(x_T, x_R, x_S) \wedge \right. \\
 & \quad \left. \neg \exists y_S \neg (\neg \text{Filme}(x_T, x_R, y_S) \vee \exists y_T \text{Filme}(y_T, \text{“Stephen Spielberg”}, y_S)) \right) \\
 & \rightsquigarrow \exists x_R \exists x_S \left(\text{Filme}(x_T, x_R, x_S) \wedge \right. \\
 & \quad \left. \neg \exists y_S (\neg \neg \text{Filme}(x_T, x_R, y_S) \wedge \neg \exists y_T \text{Filme}(y_T, \text{“Stephen Spielberg”}, y_S)) \right) \\
 & \rightsquigarrow \exists x_R \exists x_S \left(\text{Filme}(x_T, x_R, x_S) \wedge \right. \\
 & \quad \left. \neg \exists y_S (\text{Filme}(x_T, x_R, y_S) \wedge \neg \exists y_T \text{Filme}(y_T, \text{“Stephen Spielberg”}, y_S)) \right) \\
 & =: \text{SRNF}(\varphi)
 \end{aligned}$$

Folie 260

Bereichsbeschränkte Variablen einer Formel in SRNF

Definition 7.14. Per Induktion nach dem Aufbau von *SRNF-Formeln* φ definieren wir $rr(\varphi)$ folgendermaßen:

($rr(\varphi)$ steht für “range restricted variables of φ ”)

- $rr(R(v_1, \dots, v_{\text{ar}(R)})) := \{v_1, \dots, v_{\text{ar}(R)}\} \cap \text{var}$
- $rr(x=a) := rr(a=x) := \{x\}$, falls $x \in \text{var}$ und $a \in \mathbf{dom}$
- $rr(v_1=v_2) := \emptyset$, falls $\{v_1, v_2\} \subseteq \mathbf{dom}$ oder $\{v_1, v_2\} \subseteq \text{var}$
- $rr(\varphi_1 \vee \varphi_2) := rr(\varphi_1) \cap rr(\varphi_2)$
- Für $x, y \in \text{var}$ gilt: $rr(\psi \wedge x=y) := rr(x=y \wedge \psi) := \begin{cases} rr(\psi) \cup \{x, y\} & \text{falls } \{x, y\} \cap rr(\psi) \neq \emptyset \\ rr(\psi) & \text{sonst} \end{cases}$
- Ist weder φ_1 noch φ_2 von der Form $x=y$ für $x, y \in \text{var}$, so gilt: $rr(\varphi_1 \wedge \varphi_2) := rr(\varphi_1) \cup rr(\varphi_2)$
- $rr(\exists x \psi) := \begin{cases} \perp & \text{falls } x \notin rr(\psi) \\ rr(\psi) \setminus \{x\} & \text{sonst} \end{cases}$

Dabei gilt: “*Einmal* $\perp \rightsquigarrow$ *immer* \perp ”, d.h. für alle Mengen X gilt:

$$\perp = \perp \setminus X = \perp \cap X = X \cap \perp = \perp \cap \perp = \perp \cup X = X \cup \perp = \perp \cup \perp$$

- $rr(\neg\psi) := \begin{cases} \perp & \text{falls } rr(\psi) = \perp \\ \emptyset & \text{sonst} \end{cases}$

Bemerkung: Insbesondere gilt $rr(\varphi) = \perp$ oder $rr(\varphi) \subseteq \text{frei}(\varphi)$

Folie 261

Beispiele für $rr(\cdot)$

Beispiel 7.15. (a) Für die SRNf-Formel

$$\varphi := \exists x_R \exists x_S \left(\text{Filme}(x_T, x_R, x_S) \wedge \neg \exists y_S (\text{Filme}(x_T, x_R, y_S) \wedge \neg \exists y_T \text{Filme}(y_T, \text{“Stephen Spielberg”}, y_S)) \right)$$

gilt $rr(\varphi) = \{x_T\} = \text{frei}(\varphi)$.

(b) Für die SRNf-Formel

$$\psi := \neg \exists y_S \left((\exists x_R \text{Filme}(x_T, x_R, y_S)) \wedge \neg (\exists y_T \text{Filme}(y_T, \text{“Stephen Spielberg”}, y_S)) \right)$$

gilt $rr(\psi) = \emptyset \subsetneq \text{frei}(\psi) = \{x_T\}$.

($\psi = \text{SRNF}(\psi')$, wobei ψ' die allerletzte Formel von Beispiel 7.8 ist.

In der Active Domain Semantik ist ψ äquivalent zu φ)

(c) Für die SRNf-Formel

$$\chi := R(x) \wedge \exists y \left(\neg R(y) \wedge \neg \exists z (\neg R(z) \wedge \neg z=y) \right)$$

gilt $rr(\chi) = \perp$. (χ ist die Formel SRNF(“Formel aus Bsp. 7.5”))

Folie 262

Der Sichere Relationenkalkül CALC_{sr}

Definition 7.16. Sei \mathbf{S} ein Datenbankschema.

$\text{CALC}_{sr}[\mathbf{S}]$ bezeichnet die Klasse aller $\text{CALC}[\mathbf{S}]$ -Anfragen der Form $\{(e_1, \dots, e_r) : \varphi\}$, für die gilt: $rr(\text{SRNF}(\varphi)) = \text{frei}(\varphi)$.

Bemerkung:

Aus den Definitionen von $SRNF(\cdot)$ und $rr(\cdot)$ erhält man direkt einen Algorithmus, der bei Eingabe einer FO[S]-Formel φ überprüft, ob $rr(SRNF(\varphi)) = \text{frei}(\varphi)$.

Somit gibt es also auch einen Algorithmus, der bei Eingabe einer CALC[S]-Anfrage Q entscheidet, ob Q zu $CALC_{sr}[S]$ gehört oder nicht.

Ziel jetzt:

Zeige, dass alle $CALC_{sr}$ -Anfragen *bereichsunabhängig* sind und dass $CALC_{sr}$ dieselben Anfragefunktionen ausdrücken kann wie die relationale Algebra.

Folie 263

Beispiele für $CALC_{sr}$ -Anfragen

Anfragen in $CALC_{sr}$:

- Welche Filme laufen in mindestens 2 Kinos?

$$\left\{ (x_T) : \exists x_K \exists x_Z \exists y_K \exists y_Z (Programm(x_K, x_T, x_Z) \wedge Programm(y_K, x_T, y_Z) \wedge \neg x_K = y_K) \right\}$$

- In welchen Filmen hat “George Clooney” mitgespielt, aber nicht selbst Regie geführt?

$$\left\{ (x_{Titel}) : \exists x_{Regie} \left(Filme(x_{Titel}, x_{Regie}, \text{“George Clooney”}) \wedge \neg Filme(x_{Titel}, \text{“George Clooney”}, \text{“George Clooney”}) \right) \right\}$$

- Welche Filme haben nur Schauspieler, die schon mal in einem Film von “Stephen Spielberg” mitgespielt haben?

$$\left\{ (x_T) : \exists x_R \exists x_S \left(Filme(x_T, x_R, x_S) \wedge \forall y_S (Filme(x_T, x_R, y_S) \rightarrow \exists y_T Filme(y_T, \text{“Stephen Spielberg”}, y_S)) \right) \right\}$$

Nicht in $CALC_{sr}$ sind ...

- $\left\{ (x_T) : \forall y_S (\exists x_R Filme(x_T, x_R, y_S) \rightarrow \exists y_T Filme(y_T, \text{“Stephen Spielberg”}, y_S)) \right\}$
- $\left\{ (x) : R(x) \wedge \exists y (\neg R(y) \wedge \forall z (R(z) \vee z=y)) \right\}$

Folie 264

Ein technisches Lemma

Lemma 7.17. *Sei \mathbf{S} ein Datenbankschema.*

Für jede FO[\mathbf{S}]-Formel φ in SRNf mit $rr(\varphi) \neq \perp$,

für jede Datenbank $\mathbf{I} \in inst(\mathbf{S})$,

für jedes $\mathbf{d} \subseteq \mathbf{dom}$ mit $adom(\varphi, \mathbf{I}) \subseteq \mathbf{d}$ und

für jede Belegung β für φ in \mathbf{d} gilt:

(1) *Falls $\mathbf{I} \models_{\mathbf{d}} \varphi[\beta]$, so gilt für alle $x \in rr(\varphi)$: $\beta(x) \in adom(\varphi, \mathbf{I})$.*

und

(2) *Für alle $\mathbf{d}' \subseteq \mathbf{dom}$ mit $adom(\varphi, \mathbf{I}) \subseteq \mathbf{d}'$ und $\beta(\text{frei}(\varphi)) \subseteq \mathbf{d}'$ gilt:*

$$\mathbf{I} \models_{\mathbf{d}} \varphi[\beta] \iff \mathbf{I} \models_{\mathbf{d}'} \varphi[\beta]$$

Beweis: Siehe Tafel.

Folie 265

Relationenkalkül vs. Relationale Algebra

Korollar 7.18. *Die folgenden Anfragesprachen können genau dieselben Anfragefunktionen ausdrücken:*

(a) CALC_{sr}

(b) CALC_{di}

(c) CALC_{adom}

(d) *relationale Algebra (unbenannte oder benannte Perspektive)*

Und für jedes feste Datenbankschema \mathbf{S} gilt: Jede Anfrage aus einer dieser Sprachen kann in polynomieller Zeit in äquivalente Anfragen der anderen Sprachen übersetzt werden.

Beweis: Siehe Tafel.

Im weiteren Verlauf der Vorlesung schreiben wir manchmal “Relationenkalkül” oder einfach CALC, um irgendeine der Varianten CALC_{sr} , CALC_{di} bzw. CALC_{adom} des Relationenkalküls zu bezeichnen.

In der Literatur wird oft der Begriff “Relational vollständige Sprache” benutzt, um Anfragesprachen zu bezeichnen, die mindestens die Ausdrucksstärke der relationalen Algebra (bzw., äquivalent dazu, des Relationenkalküls) besitzen.

7.3 Statische Analyse und Auswertungskomplexität

Folie 266

Zur Erinnerung

Wunschliste für bessere Optimierung:

- zum “Löschen leerer Zwischenergebnisse”:
Test, ob eine gegebene (Teil-)Anfrage Q unerfüllbar ist ($“Q \equiv \emptyset”$)
- zum “Löschen von Redundanzen”:
Test, ob zwei (Teil-)Anfragen Q und P äquivalent sind ($“Q \equiv P”$)
- Wir kennen bereits: Algorithmen zum Lösen dieser Probleme für konjunktive Anfragen
- Jetzt: *Unentscheidbarkeit dieser Probleme für relationale Algebra*

Folie 267

Statische Analyse für Relational vollständige Sprachen

Satz 7.19.

Sei \mathbf{S} ein Datenbankschema, das mindestens ein Relationssymbol R enthält, dessen Stelligkeit ≥ 2 ist. Dann sind die folgenden Probleme unentscheidbar:

ERFÜLLBARKEITSPROBLEM FÜR RELATIONALE ALGEBRA ÜBER \mathbf{S}

Eingabe: Anfrage Q der relationalen Algebra über dem Datenbankschema \mathbf{S}

Frage: Ist Q erfüllbar (d.h. gibt es mind. ein $\mathbf{I} \in inst(\mathbf{S})$ mit $[[Q]](\mathbf{I}) \neq \emptyset$) ?

ÄQUIVALENZPROBLEM FÜR RELATIONALE ALGEBRA ÜBER \mathbf{S}

Eingabe: Anfragen Q und P der rel. Algebra über dem DB-Schema \mathbf{S}

Frage: Ist $Q \equiv P$ (d.h. gilt für alle $\mathbf{I} \in inst(\mathbf{S})$, dass $[[Q]](\mathbf{I}) = [[P]](\mathbf{I})$) ?

QUERY CONTAINMENT PROBLEM FÜR REL. ALGEBRA ÜBER \mathbf{S}

Eingabe: Anfragen Q und P der rel. Algebra über dem DB-Schema \mathbf{S}

Frage: Ist $Q \sqsubseteq P$ (d.h. gilt für alle $\mathbf{I} \in inst(\mathbf{S})$, dass $[[Q]](\mathbf{I}) \subseteq [[P]](\mathbf{I})$) ?

Beweis: Einfache Folgerung aus dem Satz von Trakhtenbrot (Theorem 7.11) und der Äquivalenz zwischen relationaler Algebra und dem Relationenkalkül. Details: *Übung*.

Folie 268

Auswertungsproblem: Boolesche Anfragen \rightsquigarrow beliebige Anfragen

Theorem 7.20. *Sei \mathbb{A} ein Algorithmus, der das Auswertungsproblem für Boolesche Anfragen des Relationenkalküls löst.*

Dann gibt es einen Algorithmus \mathbb{B} , der das Auswertungsproblem für (beliebige) Anfragen des Relationenkalküls unter Rückgriff auf \mathbb{A} mit Verzögerung $\mathcal{O}(k^3 \cdot n)$ löst.

Beweis: Identisch zum Beweis von Theorem 3.20.

Folgerung: Falls wir das Auswertungsproblem für *Boolesche* Anfragen des Relationenkalküls effizient lösen können, dann können wir es auch für *beliebige* Anfragen des Relationenkalküls effizient lösen.

Zur Erinnerung: Wir wissen bereits, dass das Auswertungsproblem bereits für Boolesche Anfragen des *konjunktiven Kalküls* NP-vollständig ist (Satz von Chandra und Merlin).

Folie 269

Die Komplexitätsklasse Pspace (“polynomieller Platz”)

Zur Erinnerung:

- Ein Entscheidungsproblem B gehört zur Klasse PSPACE, falls es eine (deterministische) Turingmaschine T und eine Konstante c gibt, so dass für jede Zahl N und jede zum Problem B passende Eingabe w der Größe N gilt:
Die Berechnung von T bei Eingabe w benutzt $\leq N^c$ Bandzellen und endet mit der Ausgabe “ja”, falls w eine “ja”-Instanz für B ist, bzw. mit der Ausgabe “nein”, falls w eine “nein”-Instanz für B ist.
- Es gilt: $P \subseteq NP \subseteq PSPACE = NPSPACE \subseteq EXPTIME$.
- Ein Entscheidungsproblem B heißt *PSPACE-vollständig* (bzgl. Polynomialzeit-Reduktionen), falls gilt:

- (1) $B \in \text{PSPACE}$ und
 - (2) B ist *PSPACE-hart*, d.h. für jedes Problem $A \in \text{PSPACE}$ gibt es eine *Polynomialzeit-Reduktion* f von A auf B (kurz: $f : A \leq_p B$)
- Eine *Polynomialzeit-Reduktion* f von A auf B ist eine (deterministisch) in polynomieller Zeit berechenbare Funktion, die jede zum Problem A passende Eingabe w auf eine zum Problem B passende Eingaben $f(w)$ abbildet, so dass gilt
 w ist eine “ja”-Instanz für $A \iff f(w)$ ist eine “ja”-Instanz für B .
 - Es gilt: $(A \text{ PSPACE-hart und } A \leq_p B) \implies B \text{ PSPACE-hart}$.

Folie 270

Auswertungskomplexität des Relationenkalküls

Theorem 7.21 (Chandra, Merlin, 1977). *Das Auswertungsproblem für Boolesche Anfragen des Relationenkalküls ist PSPACE-vollständig. (kombinierte Komplexität)*

Beweis: Siehe Tafel ...

Zum Nachweis der PSPACE-Härte benutzen wir das folgende Resultat (*hier ohne Beweis*):

Theorem (Stockmeyer, Meyer, 1973) *QBF ist PSPACE-vollständig.*

Das Problem QBF ist dabei folgendermaßen definiert: (*QBF steht für “Quantified Boolean Formulas”*)

QBF

Eingabe: Quantifizierte Aussagenlogische Formel ψ

Frage: Hat ψ den Wert 1 ?

“Quantifizierte Aussagenlogische Formel” bedeutet:

ψ ist von der Form $\exists X_1 \forall X_2 \exists X_3 \cdots Q_m X_m \chi$ wobei $m \geq 1$, $Q_m = \exists$, falls m ungerade und $Q_m = \forall$, falls m gerade, χ eine Aussagenlogische Formel über den Variablen X_1, \dots, X_m .

Die Formel ψ hat den Wert 1, falls gilt: Es gibt eine Belegung von X_1 (mit 0 oder 1), so dass es für jede Belegung von X_2 eine Belegung von X_3 gibt, so dass ... χ gilt.

(Andernfalls hat ψ den Wert 0.)

Folie 271

Datenkomplexität des Relationenkalküls

Datenkomplexität:

- *Blickwinkel:* Anfrage ist fest; die Eingabe besteht “nur” aus der Datenbank
- *Rechtfertigung:* i.d.R. ist die Anfrage kurz, die Datenbank aber sehr groß.
Bei DB-Anwendungen in der Praxis außerdem oft: einige (wenige) Standard-Anfragen, die immer wieder gestellt werden (d.h.: Anfragen fest, die Einträge in der Datenbank können sich mit der Zeit aber ändern).
- *Notation:* Für jede feste Anfrage Q des Relationenkalküls (bzw. der relationalen Algebra) bezeichnet EVAL_Q das folgende Auswertungsproblem:

EVAL_Q

Eingabe: Datenbank \mathbf{I} (vom zu Q “passenden” Datenbankschema)

Aufgabe: Berechne $\llbracket Q \rrbracket(\mathbf{I})$.

- Für jede feste Anfrage Q der relationalen Algebra kann EVAL_Q in *polynomieller Zeit* gelöst werden (denn gemäß Proposition 6.4 in Zeit $(k+n)^{\mathcal{O}(k)}$, wobei $k := \llbracket Q \rrbracket$)
- Dies lässt sich noch deutlich verbessern zur Aussage: EVAL_Q kann in *konstanter Zeit* gelöst werden auf *Parallelrechnern mit polynomiell vielen Prozessoren*.
Genauer: EVAL_Q gehört zur Komplexitätsklasse $\text{AC}^0 \dots$

Folie 272

Schaltkreise:

- Wir betrachten Schaltkreise mit \wedge, \vee, \neg -Gattern;
wobei \wedge und \vee -Gatter beliebig viele Eingänge haben dürfen
- Zum “Zugriff” auf die Eingabe-Datenbank \mathbf{I} außerdem:
Für jedes $R \in \mathbf{S}$, $r := \text{ar}(R)$ und alle $i_1, \dots, i_r \in \{1, \dots, |\text{adom}(\mathbf{I})|\}$
ein mögliches “Eingabe-Gatter” der Form “ $R(i_1, \dots, i_r)$ ” zum Testen,
ob das aus den entsprechenden Werten gebildete Tupel zur Relation $\mathbf{I}(R)$ gehört oder nicht.
Analog auch für jedes $c \in \text{dom}$ Eingabe-Gatter der Form “ $i=c$ ” zum Testen, ob das i -te Element in $\text{adom}(\mathbf{I})$ die Konstante c ist.

Definition 7.22 ($AC^0 \cong$ “Probleme, die mit Schaltkreisen konst. Tiefe und polyn. Größe lösbar sind”). Sei Q eine Boolesche Anfrage an Datenbanken vom Schema \mathbf{S} . Das Problem $EVAL_Q$ gehört genau dann zur *Komplexitätsklasse* AC^0 , wenn es eine Familie $(C_m)_{m \geq 1}$ von Schaltkreisen (der obigen Art) gibt, so dass gilt:

- Es gibt eine Zahl $d \in \mathbb{N}$, so dass für jedes $m \geq 1$ gilt: C_m hat die Tiefe $\leq d$.
- Es gibt eine Zahl $d' \in \mathbb{N}$, so dass für jedes $m \geq 1$ gilt: C_m besteht aus maximal $m^{d'}$ vielen Gattern.
- Für jedes $m \geq 1$ und jede Datenbank $\mathbf{I} \in inst(\mathbf{S})$ mit $|\text{adom}(\mathbf{I})| = m$ gilt:

$$C_m \text{ akzeptiert Eingabe } \mathbf{I} \iff \llbracket Q \rrbracket(\mathbf{I}) = \text{“ja”}$$

Folie 273

Theorem 7.23 (Immerman, 1987). *Für jede Boolesche Anfrage Q des Relationenkalküls gehört $EVAL_Q$ zu AC^0*

Beweis: Hier nur die Beweisidee (Details: *Übung*) anhand der Booleschen Anfrage

$$Q := \left\{ () : \underbrace{\exists x \forall y \exists z (R(x, y) \wedge \neg z=c)}_{=: \varphi} \right\}$$

Ansatz zur Konstruktion von C_m (auszuwerten über einer Datenbank \mathbf{I} mit $|\text{adom}(\mathbf{I})| = m$) :

$$\varphi \rightsquigarrow \bigvee_{i=1}^m \bigwedge_{j=1}^m \bigvee_{\ell=1}^m (R(i, j) \wedge \neg \ell=c) \rightsquigarrow \text{Schaltkreis } C_m \text{ (siehe Tafel) der Tiefe } \leq \|\varphi\|$$

7.4 Grenzen der Ausdrucksstärke

Folie 274

Beispiel: Frankfurter U-Bahn-Netz

Hier vereinfacht: Eine Relation *U-Bahn-Netz* mit Attributen *Linie*, *Halt*, *nächsterHalt*

U-Bahn-Netz

<i>Linie</i>	<i>Halt</i>	<i>nächsterHalt</i>
U4	Bockenheimer Warte	Festhalle/Messe
U4	Festhalle/Messe	Hauptbahnhof
U4	Hauptbahnhof	Willy-Brandt-Platz
U4	Willy-Brandt-Platz	Dom/Römer
...
U7
U7	Kirchplatz	Leipziger Str.
U7	Leipziger Str.	Bockenheimer Warte
U7	Bockenheimer Warte	Westend
...

Anfrage:

Gib alle Stationen aus, die von “Bockenheimer Warte” aus ohne Umsteigen zu erreichen sind.

(1) mit max. 1 Zwischenhalt:

$$\left\{ (x_S) : \exists x_L \left(U\text{-Bahn-Netz}(x_L, \text{“Bockenheimer Warte”}, x_S) \vee \exists x_Z \left(U\text{-Bahn-Netz}(x_L, \text{“Bockenheimer Warte”}, x_Z) \wedge U\text{-Bahn-Netz}(x_L, x_Z, x_S) \right) \right) \right\}$$

(2) mit max. 2 Zwischenhalten: *analog*

(3) mit *beliebig vielen* Zwischenhalten ???

Folie 275

Gaifman-Lokalität einer Anfrage

Definition 7.24. Sei **S** ein Datenbankschema, sei $r \geq 1$.

Eine Anfrage Q der Stelligkeit r heißt *Gaifman-lokal*, falls es ein $d \geq 0$ gibt, so dass für alle Datenbanken $\mathbf{I} \in \text{inst}(\mathbf{S})$ und alle Tupel $a \in \text{dom}^r$ und $b \in \text{dom}^r$ gilt:

$$\text{Falls } \mathcal{N}_d^{\mathbf{I}, Q}(a) \cong \mathcal{N}_d^{\mathbf{I}, Q}(b), \text{ so } (a \in \llbracket Q \rrbracket(\mathbf{I}) \iff b \in \llbracket Q \rrbracket(\mathbf{I})).$$

Notation hierbei:

- $\mathcal{N}_d^{\mathbf{I},Q}(a)$: die d -Nachbarschaft von $a = (a_1, \dots, a_r)$, d.h.: die Datenbank \mathbf{I} , eingeschränkt auf die Elemente $N_d^{\mathbf{I},Q}(a)$, wobei $N_0^{\mathbf{I},Q}(a) := \{a_1, \dots, a_r\} \cup \text{adom}(Q)$

$$N_{i+1}^{\mathbf{I},Q}(a) := N_i^{\mathbf{I},Q}(a) \cup \left\{ c \in \mathbf{dom} : \begin{array}{l} \text{es gibt ein Tupel } t \text{ in } \mathbf{I}, \text{ in dem so-} \\ \text{wohl } c \text{ als auch mind. ein Element} \\ \text{aus } N_i^{\mathbf{I},Q}(a) \text{ vorkommt} \end{array} \right\}$$

- $\mathcal{N}_d^{\mathbf{I},Q}(a) \cong \mathcal{N}_d^{\mathbf{I},Q}(b)$: $\mathcal{N}_d^{\mathbf{I},Q}(a)$ ist isomorph zu $\mathcal{N}_d^{\mathbf{I},Q}(b)$, d.h. es gibt eine bijektive Abbildung f von $X := N_d^{\mathbf{I},Q}(a)$ nach $Y := N_d^{\mathbf{I},Q}(b)$, so dass gilt:
 - $f(a) = b$,
 - $f(c) = c$, für alle $c \in \text{adom}(Q)$,
 - für jedes $R \in \mathbf{S}$ und jedes Tupel $t \in X^{\text{ar}(R)}$ gilt:
 $t \in \mathbf{I}(R) \iff f(t) \in \mathbf{I}(R)$.

Folie 276

Gaifman-Lokalität des Relationenkalküls

Zur Erinnerung:

Eine Anfrage Q der Stelligkeit r heißt *Gaifman-lokal*, falls es ein $d \geq 0$ gibt, so dass für alle Datenbanken $\mathbf{I} \in \text{inst}(\mathbf{S})$ und alle Tupel $a \in \mathbf{dom}^r$ und $b \in \mathbf{dom}^r$ gilt:

Falls $\mathcal{N}_d^{\mathbf{I},Q}(a) \cong \mathcal{N}_d^{\mathbf{I},Q}(b)$, so $(a \in \llbracket Q \rrbracket(\mathbf{I}) \iff b \in \llbracket Q \rrbracket(\mathbf{I}))$.

Theorem 7.25 (Gaifman-Lokalität (hier ohne Beweis)). *Jede Anfrage Q des Relationenkalküls $\text{CALC}_{\text{adom}}$ ist Gaifman-lokal.*

Beweis: Siehe Vorlesung „Logik und Komplexität“.

Folie 277

Anwendung der Gaifman-Lokalität

Beispiel 7.26. Die Anfrage “Gib alle Stationen aus, die von “Bockenheimer Warte” aus ohne Umsteigen zu erreichen sind” kann nicht im Relationenkalkül (also auch nicht in der relationalen Algebra) beschrieben werden.

Beweis: Siehe Tafel.

Bemerkung:

Mit etwas anderen Methoden kann man auch zeigen, dass der Relationenkalkül *“nicht zählen kann”*. Zum Beispiel kann die Anfrage *“Hat Stephen Spielberg mehr Filme gedreht als Alfred Hitchcock?”* nicht im Relationenkalkül ausgedrückt werden.

Beweismethode:

Ehrenfeucht-Fraïssé-Spiele; siehe Vorlesung *„Logik in der Informatik“*.

Kapitel 8

Zusammenfassung und Ausblick

8.1 Zusammenfassung

Folie 278

Ausdrucksstärke von Anfragesprachen

- *stratifiziertes Datalog*[−]
- *relational vollständige Sprachen:*
relationale Algebra, nr-Datalog[−],
Relationenkalkül (Varianten: active domain, bereichsunabhängig,
safe-range)
- *Positive Anfragen:*
SPCU, SPJRU, nr-Datalog
positiver existentieller Kalkül PE-CALC_{adom}
- *konjunktive Anfragen:*
SPC, SPJR, regelbasierte konjunktive Anfragen, konjunktiver Kalkül,
Tableau-Anfragen
- *azyklische konjunktive Anfragen:*
azyklische regelbasierte konjunktive Anfragen, Semijoin-Anfragen,
konjunktives Guarded Fragment
- *Datalog*

Schematische Darstellung: siehe Tafel

Folie 279

Methoden zum Nachweis von Grenzen der Ausdrucksstärke einer Anfragesprache

- Monotonie von regelbasierten konjunktiven Anfragen und Datalog-Anfragen
- regelbasierte konjunktive Anfragen und Datalog-Anfragen Q sind abgeschlossen unter $\text{adom}(Q)$ -Homomorphismen
- Gaifman-Lokalität von Anfragen des Relationenkalküls

Folie 280

Auswertungskomplexität und Statische Analyse

	Datenkomplexität	kombinierte Komplexität	Erfüllbarkeitsproblem	Äquivalenzproblem	Query Containment Problem
Stratifiziertes Datalog ⁻	P- <i>vollständig</i>	EXPTIME- <i>vollständig</i>	unentscheidbar	unentscheidbar	unentscheidbar
Relationale Algebra	in AC^0	PSPACE- <i>vollständig</i>	unentscheidbar	unentscheidbar	unentscheidbar
Positive Anfragen	in AC^0	NP- <i>vollständig</i>	entscheidbar	<i>entscheidbar</i>	<i>entscheidbar</i>
Konjunktive Anfragen	in AC^0	NP- <i>vollständig</i>	in P	NP- <i>vollständig</i>	NP- <i>vollständig</i>
Azykl. Konj. Anfragen	in AC^0	in P (LogCFL- <i>vollst.</i>)	in P	in P	in P
Datalog	P- <i>vollständig</i>	EXPTIME- <i>vollständig</i>	entscheidbar	unentscheidbar	unentscheidbar

Folie 281

Wichtige Stichpunkte

- Homomorphismus-Satz
- Algorithmus zur Minimierung konjunktiver Anfragen
- Sätze von Chandra und Merlin
- Satz von Trakhtenbrot (und Folgerungen daraus)
- Satz von Knaster und Tarski

Folie 282

Funktionale Abhängigkeiten

- The Chase („Die Verfolgungsjagd“)
- Äquivalenz, Query Containment und Minimierung konjunktiver Anfragen
unter Berücksichtigung funktionaler Abhängigkeiten
- effizienter Test, ob $\mathcal{F} \models f$, bei Eingabe einer FD-Menge \mathcal{F} und einer FD f
- Armstrong-Kalkül

8.2 Ausblick auf weitere Themen

Folie 283

Einige weiterführende Themen

- Semistrukturierte Daten und XML
- Datenbanken mit unvollständiger Information
- Datenaustausch und Datenintegration
- Datenströme
- Constraint Datenbanken
- Probabilistische Datenbanken

Folie 284

– *ENDE* –