IMDb (Teil 2)
Relationale Algebra

VL Big Data Engineering
(vormals Informationssysteme)

Prof. Dr. Jens Dittrich

bigdata.uni-saarland.de

5. Mai 2022

IMDb (Teil 2)

1. Konkrete Anwendung: IMDb

- Das haben wir schon letztes Mal angeschaut.

IMDb (Teil 2)

2. Was sind die Datenmanagement und -analyseprobleme dahinter?

letzte Woche:

Frage 1

Wie werden in IMDb die Daten zu Filmen, Schauspielern, Regisseuren usw. modelliert und abgelegt?

Frage 2

Wie werden in IMDb Verknüpfungen dieser Daten modelliert und abgelegt?

diese Woche:

Frage 3

Wie stellen wir Anfragen an diese Daten?

...

IMDb (Teil 2)

3. Grundlagen, um diese Probleme lösen zu können
   (a) Folien
   (b) Jupyter/Python/SQL Hands-on

- Die relationale Algebra

Wichtigste Lernziele

Relationale Algebra

Operatoren (unär und binär), Selektion (nicht zu verwechseln mit SELECT in SQL), Projektion, Vereinigung, Differenz, Kreuzprodukt, Schnitt, Theta Join, Gruppierung vs Aggregation

Die relationale Algebra: Operatoren

- modulare Anfragesprache zur Umwandlung einer oder mehrerer Eingaberelationen in eine Ausgaberelation
- besteht aus einer Menge von Operatoren:

  unär: op: $\mathcal{P}(R) \to \mathcal{P}(R)$

  binär: op: $\mathcal{P}(R) \times \mathcal{P}(R) \to \mathcal{P}(R)$

  hierbei bezeichnet $\mathcal{P}(R)$ die Menge aller möglichen Relationen

- aus historischen Gründen nur unäre und binäre Operatoren (dies ist eigentlich eine unnötige Einschränkung)

Achtung

Die Relationale Algebra beschreibt nur abstrakt **WAS** berechnet werden soll, aber nicht **WIE** diese Berechnung algorithmisch umgesetzt wird!

Stellenwert für Datenanalyse und -verarbeitung:

Vergleichbar mit dem Periodensystem der Elemente in der Chemie

Die relationale Algebra

- gibt es in verschiedenen Varianten/Dialekten
- der Kern der Operatoren (wie im Folgenden erklärt) ist aber immer gleich
- kann sehr leicht erweitert werden
- es gibt viele verschiedene Implementierungen der relationalen Algebra, die aktuell Bekannteste ist Apache Spark.
- wird intern von Systemen zur Anfrageoptimierung verwendet, z.B. in relationalen Datenbanksystemen
- oder: wird vom Nutzer spezifiziert als Datenflussprogramm, z.B. in Apache Spark

Anmerkungen zu Domänen

Wir hatten bereits für Entity-Relationship-Modellierung Domänen (Wertebereiche) eingeführt (siehe IMDb 01, Folie 19):

Im Folgenden gilt:

Domäne (Wertebereich)

Eine Domäne ist eine Menge **atomarer** Werte. Diese Werte dürfen nicht strukturiert (d.h. weiter aufteilbar) sein. Domänen werden notiert als D, z.B. integer, float, String, etc.

Vorsicht

Diese Einschränkung ist veraltet und führt zu sehr viel Verwirrung, z.B. der sogenannten ersten Normalform. In der Praxis wird diese Einschränkung meist abgeschwächt. Wir werden darauf zurückkommen, wenn wir (modernes) SQL-99 diskutieren.

Welche atomaren Domänen sind erlaubt?

Aber was für atomare Domänen sind **konkret** erlaubt in ER, im Relationalen Modell, in der Relationalen Algebra, ...?

Erlaubte Domänen in Werkzeug/Modellierungstechnik X

Die erlaubten Domänen werden meist durch das Werkzeug vorgegeben und dann näher spezifiziert. Falls nicht näher spezifiziert nehmen wir kanonische Wertebereiche an wie int/integer, float, string/str/varchar/text, bool/boolean. Die Details dieser Domänen (insbesondere Überlauf) spielen im Moment noch keine Rolle.

Das Periodensystem der Relationalen Algebra

Symbol	Name		Einteilung	
	Deutsch	Englisch	Klasse	unär/binär
σ	Selektion	selection	Basisoperatoren	unär
π	Projektion	projection		unär
$\cup$	Vereinigung	union		binär
$-$	Differenz	minus		binär
$\times$	kartesisches Produkt (oder Kreuzprodukt)	cartesian product		binär
ρ	Umbenennung	rename		unär
$\cap$	Schnitt	intersection	abgeleitete Operatoren	binär
$\bowtie$	Verbund	join		
$\bowtie_\theta$	Theta-Verbund	theta join		
$\bowtie_{[L],[R]}$	Equi-Verbund	equi join		
$\ltimes$	Linker Pseudo-Verbund	left semi join		
$\rtimes$	Rechter Pseudo-Verbund	right semi join		
$\triangleright$	Linker Anti-Pseudo Verbund	left anti semi join		
$\triangleleft$	Rechter Anti-Pseudo Verbund	right anti semi join		
$⟕$	Linker äußerer Verbund	left outer join		
$⟖$	Rechter äußerer Verbund	right outer join		
$⟗$	Äußerer Verbund	full outer join		
γ	Gruppierung (mit Aggregation)	grouping (group by)	Erweiterungen	unär
Γ	Co-Gruppierung (**ohne** Aggregation)	co-grouping		binär
	...			

Anmerkung zum relationalen Modell

Da meist aus dem Kontext klar ist, ob bei einer Relation $R = (\{R\}, [R])$, die Relation R oder die Ausprägung $\{R\}$ gemeint ist, schreiben wir im Folgenden oft R statt $\{R\}$.

Basisoperatoren: Selektion σ

Die nachfolgenden Operatoren heißen die **Basisoperatoren** der relationalen Algebra. Viele andere Operatoren bauen hierauf auf bzw. sind *syntaktischer Zucker (syntactic sugar)*.

Selektion (selection/filter) σ

Die Selektion selektiert aus der Eingaberelation R eine Teilmenge $R' \subseteq R$ anhand eines Prädikats $P : [R] \rightarrow bool$. Das Prädikat muss wohldefiniert sein auf $[R]$.
$R' = \sigma_P(R) = \{r \in R \mid P(r)\}$.

Beispiele:
$[R_1]$: {[id:int]}, $R_1 = \{(1), (2)\}$,
$[R_2]$: {[id:int]}, $R_2 = \{(2), (3), (4)\}$

P:= id≤ 2

$\sigma_P(R_1) = \{(1), (2)\}$, $\sigma_P(R_2) = \{(2)\}$

directors		
id	**first_name**	**last_name**
78273	Quentin	Tarantino
43095	Stanley	Kubrick
11652	James (I)	Cameron

$\sigma_{id>45000}$(directors):

id	first_name	last_name
78273	Quentin	Tarantino

$\sigma_{first\_name='Stanley'}$(directors):

id	first_name	last_name
43095	Stanley	Kubrick

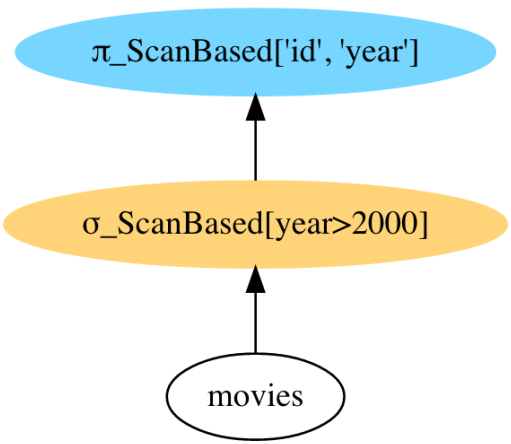
Prädikate

Prädikate

Ein Prädikat ist eine Funktion $P : [R] \rightarrow \mathit{bool}$. Es ist wohldefiniert auf dem Schema $[R]$, wenn alle in P genutzten Attribute in $[R]$ mit passenden Domänen definiert sind.
Atomare Prädikate haben die Form $A_1 = A_2$ oder $A_1 = c$, wobei A_1, A_2 Attribute und c eine Konstante ist. Die atomaren Prädikate können mit boolschen Operatoren verknüpft werden.
Zusätzlich gibt es für spezifische Domänen die folgenden Operatoren: $\leq, \geq, <, >$ (int, float) und is_substr (string). Diese bilden Vergleiche und Teilstringeigenschaft ab.

Beispiele (siehe Notebook „Relational Algebra“)

```
In [11]: exp2 = Projection_ScanBased(newmovies, ['id', 'year'])

In [12]: print(exp2)

π_ScanBased_['id', 'year'](σ_ScanBased_[year>2000](movies))

In [13]: exp2.evaluate().print_set()

[Result] : {[ id:int, year:int ]}
{
        (96779, 2001),
        (393538, 2003),
        (176712, 2004),
        (176711, 2003),
        (127297, 2003),
        (159665, 2006),
        (105938, 2002),
        (10934, 2005)
}

In [14]: graph = exp2.get_graph()
         Source(graph)

Out[14]:
```



github: Relational Algebra.ipynb

Ausdrücke auf Relationenschemata

Da Relationenschemata keine Mengen sind, benötigen wir Definitionen, um Aussagen über sie zu treffen.

Ausdrücke auf Relationenschemata

Attribut ist enthalten:
$A \in \{[A_1, \ldots, A_n]\} \Leftrightarrow$ es gibt ein j mit $1 \leq j \leq n$ und $A_j = A$.

Teilschema:
$[R] \subseteq [S]$, $\Leftrightarrow$ für alle $A \in [R]$ gilt $A \in [S]$.

Schnitt zweier Schemata (gemeinsame Attribute):
$[R] \cap [S] = [T]$ mit $A \in [T] \Leftrightarrow A \in [R] \wedge A \in [S]$.

Vereinigung zweier Schemata:
$[R] \cup [S] = [T]$ mit $A \in [T] \Leftrightarrow A \in [R] \vee A \in [S]$.

Konkateniertes Schema:
$[R] \circ [S] = [R] \cup [S]$ falls $[R] \cap [S] = \varnothing$, sonst nicht definiert

Projektion π

Projektion (projection) π

Die Projektion projiziert die Eingaberelation R auf eine Teilmenge der Attribute der Eingaberelation.
Sei $[R'] \subseteq [R]$ eine beliebige nichtleere Teilmenge der Attribute von R.
Dann ist das Ergebnis der Projektion $R' := \pi_{[R']}(R) := \{r_{[R']} \mid r \in R\}$.
$r_{[R']}$ ist ein Tupel, das nur die Attribute enthält, die in $[R']$ enthalten sind.

Syntaktischer Zucker: $\pi_{A_1,\ldots,A_n}(R) = \pi_{\{[A_1:D_1,\ldots,A_n:D_n]\}}(R)$

Beispiele:

[R] : {[a:int, b:int]}, R = {(1,3), (2,4), (2,7), (3,3), (4,2)}

$[R'] = \{[a:int]\}$

$\pi_{[R']}(R) = \pi_a(R) = \{ (1), (2), (3), (4) \}$

$[R''] = \{[b:int]\}$

$\pi_{[R'']}(R) = \pi_b(R) = \{ (3), (4), (7), (2) \}$

Projektion und Duplikate

Beachte: Projektion und Duplikate

Es gilt, dass $|\pi_{[R']}(R)| \leq |R|$.

Warum?

Durch die Projektion können Duplikate entstehen. Da eine Relation eine Menge ist, werden die Duplikate nicht mehrfach repräsentiert.

Beispiel:

genre
Comedy
Mystery
Action
Romance
Fantasy
Horror
Music
Film-Noir
Sci-Fi
Drama
Short
Documentary
Adventure
Crime
War
Family
Thriller

$|\pi_{\text{genre}}(\text{movies\_genres})| = 17$

aber:

$|\text{movies\_genres}| = 102$

Vereinigung ∪

Vereinigung (union) ∪:

Die Vereinigung vereinigt die Tupel aus zwei Relationen R_1 und R_2 zu einer neuen Relation R'. Es muss gelten: $[R_1] = [R_2]$. Dann ist das Ergebnis der Vereinigung $R' := R_1 \cup R_2$.
Die Anzahl der Tupel in R' ist beschränkt durch $max(|R_1|, |R_2|) \leq |R'| \leq |R_1| + |R_2|$.

Beispiel: sowohl neue als auch gute Filme in einer Relation

$\sigma_{year>2000}(movies) \cup \sigma_{rank \geq 7.5}(movies)$

id	name	year	rank
10934	Aliens of the Deep	2005	6.5
92616	Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb	1964	8.7
105938	Expedition: Bismarck	2002	7.5
387728	ER	1994	7.7
96779	Earthship.TV	2001	5.6
393538	Jimmy Kimmel Live!	2003	6.7
267038	Pulp Fiction	1994	8.7
121538	Full Metal Jacket	1987	8.2
193519	Lolita	1962	7.6
127297	Ghosts of the Abyss	2003	6.7
328285	Terminator, The	1984	7.9
..	..	..	..

Differenz –

Differenz (minus/except) –

Die Differenz entfernt die Tupel der Relation R_2 aus Relation R_1. Es muss gelten: $[R_1] = [R_2]$. Dann ist das Ergebnis der Differenz $R' := R_1 - R_2$.
Die Anzahl der Tupel in R' ist beschränkt durch $0 \leq |R'| \leq |R_1|$.

Beispiel:

nur neue Filme, die kein schlechtes Ranking haben, in einer Relation:
mit anderen Worten: nur neue, gute Filme

$\sigma_{\text{year}>2000}(\text{movies}) - \sigma_{\text{rank}<7.5}(\text{movies})$

id	name	year	rank
105938	Expedition: Bismarck	2002	7.5
176711	Kill Bill: Vol. 1	2003	8.4
176712	Kill Bill: Vol. 2	2004	8.2
159665	Inglorious Bastards	2006	8.3

Kreuzprodukt $\times$

Kreuzprodukt (cross product, cross join) $\times$

Das Kreuzprodukt bildet das kartesische Produkt[1] von zwei Relationen. D.h. jedes Tupel aus der linken Eingabe wird mit jedem Tupel aus der rechten Relation zu einem neuen Tupel kombiniert.
Falls $[R_1] \cap [R_2] \neq \varnothing$, siehe Eindeutigkeit von Attributnamen. Es gilt: $[R'] = [R_1] \circ [R_2]$. Das Ergebnis des Kreuzproduktes ist $R' := R_1 \times R_2 = \{t_1 \circ t_2 \mid t_1 \in R_1 \wedge t_2 \in R_2\}$, wobei $\circ$ die Konkatenation zweier Tupel ist. Die Anzahl der Tupel im Kreuzprodukt ist $|R'| = |R_1 \times R_2| = |R_1| \cdot |R_2|$.

Beispiel:
$[R_1]$: {[a:int]}, $R_1 = \{(1), (2)\}$,
$[R_2]$: {[b:int]}, $R_2 = \{(2), (3), (4)\}$

$R_1 \times R_2 = \{ (1, 2), (1,3), (1,4), (2, 2), (2,3), (2,4)\}$

$|R_1| = 2$, $|R_2| = 3$, $|R_1 \times R_2| = |R_1| \cdot |R_2| = 2 \cdot 3 = 6$. $[R_1] \cap [R_2] = \varnothing$

[1] Die Begriffe Kartesisches- und Kreuzprodukt werden für die relationale Algebra synonym verwendet.

Umbenennung ρ

Umbenennung (rename) ρ

Die Umbenennung ändert den Namen einer Relation oder den Namen eines Attributs einer Relation. Für das Umbenennen einer Relation von R zu R_1 gilt $[R'] = [R_1], \{R'\} = \{R_1\}$ und somit $R' := \rho_{R'}(R) = (\{\text{R'}\}, [\text{R'}])$. Für das Umbenennen eines Attributs $\text{A} \in [R]$ zu A' gilt: $R' := \rho_{\text{A}' \leftarrow \text{A}}(R)$. Das A in $[R]$ wird dann in A' umbenannt.

Beispiele:

siehe Python-Notebook

Das Problem mit doppelten Attributnamen

Beispiel:

$[R_1] : \{[\text{id:int}]\}$, $R_1 = \{(1), (2)\}$,
$[R_2] : \{[\text{id:int}]\}$, $R_2 = \{(3), (4)\}$
$\Rightarrow [R_1 \times R_2] = \{[\text{id:int, id:int}]\}$, $R_1 \times R_2 = \{(1,3), (1,4), (2,3), (2,4)\}$

- Wie referenzieren wir im Ergebnis des Kreuzproduktes das Attribut ‘id‘?
- Das ist nicht eindeutig! Und deswegen auch nicht erlaubt
- Deswegen muss man hier die Attribute vorher umbenennen

Abgeleitete Operatoren: Intersection ∩

Alle nachfolgenden Operatoren können auch durch die Basisoperatoren ausgedrückt werden und sind folglich nur „syntactic sugar“.

Schnitt (intersection) ∩

Der Schnitt bildet die Schnittmenge der Tupel aus zwei Relationen R_1 und R_2. Es muss gelten: $[R_1] = [R_2]$. Dann ist das Ergebnis des Schnitts $R' := R_1 \cap R_2 = R_1 - (R_1 - R_2)$.

Beispiel:

nur neue gute Filme in einer Relation:

$\sigma_{\text{year}>2000}(\text{movies}) \cap \sigma_{\text{rank}\geq 7.5}(\text{movies})$

id	name	year	rank
176711	Kill Bill: Vol. 1	2003	8.4
105938	Expedition: Bismarck	2002	7.5
176712	Kill Bill: Vol. 2	2004	8.2
159665	Inglorious Bastards	2006	8.3

Theta Join $\bowtie_\theta$

Theta-Verbund (theta join) $\bowtie_\theta$

Der Theta-Verbund bildet den Verbund aus zwei Relationen R_1 und R_2 unter dem allgemeinen Join-Prädikat θ. Hieraus wird eine neue Relation TJ erzeugt. Es gilt: $[TJ] = [R_1] \circ [R_2]$. $\theta : [TJ] \to$ bool.
$TJ := R_1 \bowtie_\theta R_2 = \sigma_\theta(R_1 \times R_2) \subseteq R_1 \times R_2$. Die Anzahl der Tupel im Joinergebnis ist $0 \leq |TJ| \leq |R_1 \times R_2| = |R_1| \cdot |R_2|$.
Falls $[R_1] \cap [R_2] \neq \emptyset$, gelten dieselben Regeln wie für die Eindeutigkeit von Attributnamen beim Kreuzprodukt.

Beispiel:
[R] : {[a:int, b:int]}, R = {(3,1), (4,2), (7,2), (3,3), (7,6)}
[S] : {[c:int, d:int]}, S ={(2,3), (1,4), (5,4), (3,8), (2,5)}

[TJ] = [R] ∪ [S] = {[a:int, b:int, c:int, d:int]}

$R \bowtie_{b=c} S = \{ (3,\underline{1},\underline{1},4), (4,\underline{2},\underline{2},3), (4,\underline{2},\underline{2},5), (7,\underline{2},\underline{2},3), (7,\underline{2},\underline{2},5), (3,\underline{3},\underline{3},8) \}$

$R \bowtie_{a=d} S = \{ (\underline{3},1,2,\underline{3}), (\underline{4},2,1,\underline{4}), (\underline{4},2,5,\underline{4}), (\underline{3},3,2,\underline{3}) \}$

Theta Join-Beispiel für IMDb

Beispiel:

directors $\bowtie_{\text{id=director\_id}}$ movies_directors

id	first_name	last_name	director_id	movie_id
43095	Stanley	Kubrick	43095	176891
43095	Stanley	Kubrick	43095	177019
43095	Stanley	Kubrick	43095	250612
11652	James (I)	Cameron	11652	328277
43095	Stanley	Kubrick	43095	30431
43095	Stanley	Kubrick	43095	106666
78273	Quentin	Tarantino	78273	267038
43095	Stanley	Kubrick	43095	65764
78273	Quentin	Tarantino	78273	118367
..	..	..	..	..

Equi Join

Equi-Verbund (equi join)

Der Equi-Verbund bildet den Verbund aus zwei Relationen R_1 und R_2 unter einem Equi-Join-Prädikat. D.h. anstatt das Join-Prädikat direkt zu spezifizieren, wie beim Theta-Join, werden Teilmengen der Schemata von R_1 und R_2 festgelegt, deren Attributwerte gleich sein müssen.

Es seien $[R_1'] \subseteq [R_1]$ sowie $[R_2'] \subseteq [R_2]$, $|[R_1']| = |[R_2']|$.

Dann wird hierdurch folgendes Joinprädikat definiert:
$\theta(r_1 \circ r_2) := \pi_{R_1'}(r_1 \circ r_2) == \pi_{R_2'}(r_1 \circ r_2)$.

Beispiel:
folgende Teilschemata wurden festgelegt:
$[\mathrm{R}_1'] = \{[\mathrm{ID}, \mathrm{Gehalt}]\}$ und $[\mathrm{R}_2'] = \{[\mathrm{SID}, \mathrm{Bonus}]\}$

$\Rightarrow \theta(r_1 \circ r_2) : \mathrm{ID} == \mathrm{SID} \wedge \mathrm{Gehalt} == \mathrm{Bonus}.$

$\mathrm{R}_1 \bowtie_{[\mathrm{R}_1'],[\mathrm{R}_2']} \mathrm{R}_2 = \mathrm{R}_1 \bowtie_{\{[\mathrm{ID},\mathrm{Gehalt}]\},\{[\mathrm{SID},\mathrm{Bonus}]\}} \mathrm{R}_2 = \mathrm{R}_1 \bowtie_\theta \mathrm{R}_2$

Equi Join: Formal

Equi Join

$R_1 \bowtie_{\{[A_1:D_1,\ldots,A_n:D_n]\},\{[B_1:D_1,\ldots,B_n:D_n]\}} R_2 =$
$R_1 \bowtie_{A_1=B_1 \wedge \ldots \wedge A_n=B_n} R_2$

Syntaktischer Zucker:
$R_1 \bowtie_{A_1,\ldots,A_n;B_1,\ldots,B_n} R_2 =$
$R_1 \bowtie_{\{[A_1,\ldots,A_n]\},\{[B_1,\ldots,B_n]\}} R_2 =$
$R_1 \bowtie_{\{[A_1:D_1,\ldots,A_n:D_n]\},\{[B_1:D_1,\ldots,B_n:D_n]\}} R_2$

Erweiterungen: Gruppierung γ

Gruppierung (group by) γ

Die Gruppierung gruppiert eine Eingaberelation anhand einer (möglicherweise leeren) Teilmenge von Attributen und erzeugt ein Tupel für jede so entstandene Gruppe (durch **Aggregation**). Seien $[G] \subseteq [R]$ und $[G_1] \subseteq [R], \ldots, [G_n] \subseteq [R]$ beliebige (auch leere) Teilmengen der Attribute der Relation R. Ferner seien Aggregatfunktionen definiert mit $f_1 : [G_1] \to [D_1], \ldots, f_n : [G_n] \to [D_n]$.
Dann ist das Ergebnis der Gruppierung $R' := \gamma_{[G],f_1([G_1]),\ldots,f_n([G_n])}(R)$.

Syntaktischer Zucker: Schemata werden abgekürzt wie bei der Projektion.

Beispiele:

$\gamma_{\text{gender, count(*)}}(\text{actors})$

gender	count(*)
F	289
M	802

$\gamma_{\text{first\_name, last\_name, count(*)}}(\text{directors} \bowtie_{\text{id=director\_id}} \text{movies\_directors})$

last_name	first_name	count(*)
Tarantino	Quentin	10
Kubrick	Stanley	16
Cameron	James (I)	14

Die wichtigsten Aggregatfunktionen

sum(A)

Berechnet die Summe der Werte von Attribut A.

avg(A)

Berechnet den Durchschnitt (arithmetisches Mittel) der Werte von Attribut A.

min(A)

Berechnet das Minimum der Werte von Attribut A.

max(A)

Berechnet das Maximum der Werte von Attribut A.

count(*)

Berechnet die Anzahl der Tupel (unabhängig von einem Attribut).

Gruppierung

Der Name „Gruppierung“ ist für diesen Operator eigentlich irreführend. Denn bei der „Gruppierung“ werden drei verschiedene Operationen durchführt:

1. **Gruppierung:** es werden alle Tupel der Eingaberelation nach den Attributen in $[G]$ gruppiert (mit anderen Worten: **horizontal partitioniert**). Alle Tupel, die bezüglich $[G]$ die gleichen Werte haben, kommen in dieselbe Gruppe/Horizontale Partition.
2. **Aggregation:** für jede in (1.) entstandene Gruppe/Horizontale Partition werden Aggregatfunktionen $\mathrm{f}_1([G_1]), \ldots, \mathrm{f}_n([G_n])$ berechnet. D.h. **für jede Gruppe/Horizontale Partition wird jede dieser Funktionen unabhängig** berechnet.
3. **Projektion:** für jede in (1.) entstandene Gruppe/Horizontale Partition wird ein Tupel mit den zugehörigen Aggregaten aus (2.) erzeugt. Das Schema dieses Tupels ist $[G] \circ [\mathrm{D}_1, \ldots, \mathrm{D}_n]$.

Gruppierung: Formal (1/2)

Horizontale Partition(ierung) (Definition mittels Gleichheit der Gruppierungsschlüssel)

Sei $[G] = \{[A_1 : D_1, \ldots, A_n : D_n]\} \subseteq [R]$.
Sei für jedes Tupel $t = (a_1, \ldots, a_n) \in \pi_{[G]}(R)$ die Relation R_t wie folgt definiert: $R_t = \sigma_{A_1=a_1 \wedge \ldots \wedge A_n=a_n}(R)$

Dann ist R_t eine Horizontale Partition (kurz HP).
Die Menge $HPT_{[G]}(R) = \left\{R_t \mid t \in \pi_{[G]}(R)\right\}$ heißt eine Horizontale Partitionierung (kurz HPT) von R.

Es gilt:
(1.) $R_i \cap R_j = \varnothing \; \forall_{i,j \in \pi_{[G]}(R) \wedge i \neq j}$ (Disjunktheit)
(2.) $\bigcup_{t \in \pi_{[G]}(R)} R_t = R$ (Vollständigkeit)

Gruppierung: Formal (2/2)

Gruppierung

Sei $[G] = \{[A_1 : D_1, \ldots, A_n : D_n]\} \subseteq [R]$.

Sei $GB := \gamma_{[G],\mathrm{f}_1([G_1]),\ldots,\mathrm{f}_m([G_m])}(R)$

Dann ist $GB := \left\{ t \circ f_1\left(\pi_{[G_1]}(R_t)\right) \circ \ldots \circ f_m\left(\pi_{[G_m]}(R_t)\right) \mid R_t \in HPT_{[G]}(R) \right\}$.

$[GB] = [G] \circ [\mathrm{D}_1, \ldots, \mathrm{D}_n]$

Horizontale Partitionierung (allgemein)

Horizontal Partitioning (allgemein)

Sei R eine Relation. Jede Zuordnung von Tupeln aus R zu Relationen $R_1, \ldots, R_k$ heißt *Horizontale Partitionierung von* R (HPT), falls $\forall_{t \in R} \exists R_i, 1 \leq i \leq k$ mit $t \in R_i$. Die Relationen R_i heißen die *Horizontalen Partitionen* (HP) von R.

Beispiele:

$R = \{(2, A), (7, B), (1, B), (6, C)\}$

- $R_1 = \{(2, A), (1, B)\}$, $R_2 = \{(7, B), (6, C)\}$ ist eine HPT.
- $R_1 = \{(1, B)\}$, $R_2 = \{(7, B), (2, A), (6, C)\}$ ist eine HPT.
- $R_1 = \{(2, A), (1, B)\}$, $R_2 = \{(2, A), (6, C)\}$ ist **keine** HPT.

Disjunkte Horizontale Partitionierung

Eine Horizontale Partitionierung heißt *disjunkt*, falls $R_i \cap R_j = \varnothing \ \forall i, j \neq i$.

$R_1 = \{(2, A), (1, B)$, $R_2 = \{(7, B), (2, A), (6, C)\}$ ist eine HPT aber **nicht** disjunkt.

Partitionierungsfunktion

Partitionierungsfunktion

Sei D eine beliebige Domäne. Jede Funktion $p : [R] \to D$ heißt *Partitionierungsfunktion* von R.

Beispiele:

$[R] = \{[a : int, b : char]\}, R = \{(2, A), (7, B), (1, B), (6, C)\}$

$p_0 : [R] \to int, p_0(t) := t.a \text{ modulo } 2$

- $p_0((2, A)) = 0$
- $p_0((7, B)) = 1$
- $p_0((1, B)) = 1$
- $p_0((6, C)) = 0$

$p_1 : [R] \to char, p_1(t) := t.b$

- $p_1((2, A)) = A$
- $p_1((7, B)) = B$
- $p_1((1, B)) = B$
- $p_1((6, C)) = C$

Induzierte Horizontale Partitionierung

Induzierte Horizontale Partitionierung

Sei $p : [R] \to D$ eine Partitionierungsfunktion. Die Horizontale Partitionierung von R in Partitionen R_i so dass $\forall_{t \in R} t \in R_{p(t)}$ heißt *Induzierte Horizontale Partitionierung*.

Beispiele:

$p_0 : [R] \to int, p_0(t) := t.a$ modulo 2

- $p_0((2, A)) = 0$
- $p_0((7, B)) = 1$
- $p_0((1, B)) = 1$
- $p_0((6, C)) = 0$

Induzierte Horizontale Partitionierung:
$R_0 = \{(2, A), (6, C)\}$,
$R_1 = \{(7, B), (1, B)\}$

$p_1 : [R] \to char, p_1(t) := t.b$

- $p_1((2, A)) = A$
- $p_1((7, B)) = B$
- $p_1((1, B)) = B$
- $p_1((6, C)) = C$

Induzierte Horizontale Partitionierung:
$R_A = \{(2, A)\}$,
$R_B = \{(7, B), (1, B)\}$,
$R_C = \{(6, C)\}$

Co-Gruppierung Γ: Intuition

Horizontale Co-Partition(ierung), aka Co-Gruppierung Γ

Die Co-Gruppierung partitioniert zwei Eingaberelationen R und S anhand zweier Partitionierungsfunktionen $p_R : [R] \to D$ und $p_S : [S] \to D$, die dieselbe Zieldomäne D haben. Die Ausgabe enthält ein Tupel für jedes Tupel in $\left\{ d \mid d \in p_R(r) \forall r \in R \ \vee \ d \in p_S(s) \forall s \in S \right\}$.
Die Ausgabe der Co-Gruppierung hat das Schema $\left[\text{d:D, R:[R], S:[S]}\right]$.

Beispiel:
[R] : {[a:int, b:int]}, R = {(3,1), (4,2), (7,2), (3,3), (7,6)}
[S] : {[c:int, d:int]}, S ={(2,3), (1,4), (5,4), (3,8), (2,5)}

$p_R : [R] \to int, p_R(r) := r.b,\ p_S : [S] \to int, p_S(s) := s.c$

$$[\text{HCPT}]_{p_R,p_S} = \Big\{ \underbrace{\big(1, \{(3,1)\}, \{(1,4)\}\big)}_{d=1}, \underbrace{\big(2, \{(4,2),(7,2)\}, \{(2,3),(2,5)\}\big)}_{d=2},$$
$$\underbrace{\big(3, \{(3,3)\}, \{(3,8)\}\big)}_{d=3}, \underbrace{\big(6, \{(7,6)\}, \{\}\big)}_{d=6}, \underbrace{\big(5, \{\}, \{(5,4)\}\big)}_{d=5} \Big\}$$

Co-Gruppierung Γ: Formal

Horizontale Co-Partition(ierung), aka Co-Gruppierung Γ

Seien:
$HPT_{p_R}(R) = \left\{R_{p_R(t)} \mid t \in R\right\}$ eine HPT von R und
$HPT_{p_S}(S) = \left\{S_{p_S(t)} \mid t \in S\right\}$ eine HPT von S.

Dann ist:
$HCPT_{p_R,p_S}(R,S) = \left\{(d, R_d, S_d) \mid R_d \in HPT_{p_R}(R), S_d \in HPT_{p_S}(S))\right\}$
die Co-Gruppierung (kurz HCPT) von R und S.

Zusammenhang von Co-Gruppierung und Equi-Join: Formal

Äquivalenz von Co-Gruppierung (plus Kreuzprodukt) und Equi-Join

Sei:
$HCPT_{p_S,p_T}(S,T)$ eine HCPT von S und T.

$[S' \subseteq S], [T' \subseteq T], |[S']| = |[T']|$ seien die gewünschten Teilschemata eines Equi-Join.

$p_S : [S] \to D, p_S(s) := \pi_{[S']}(s)$
$p_T : [T] \to D, p_T(t) := \pi_{[T']}(t)$

Dann gilt:
$\left\{ v \mid v \in S_d \times T_d \text{ mit } (d, S_d, T_d) \in HCPT_{p_S,p_T}(S,T) \right\} = S \bowtie_{[S'],[T']} T.$

Beispiel

[R] : {[a:int, b:int]}, R = {(3,1), (4,2), (7,2), (3,3), (7,6)}
[S] : {[c:int, d:int]}, S ={(2,3), (1,4), (5,4), (3,8), (2,5)}

$p_R : [R] \to int, p_R(r) := r.b,\ p_S : [S] \to int, p_S(s) := s.c$

$$[\text{HCPT}]_{p_R,p_S} = \Big\{ \underbrace{\big(1, \{(3,1)\}, \{(1,4)\}\big)}_{d=1,\{(3,\underline{1},\underline{1},4)\}}, \underbrace{\big(2, \{(4,2),(7,2)\}, \{(2,3),(2,5)\}\big)}_{d=2,\{(4,\underline{2},\underline{2},3),(4,\underline{2},\underline{2},5),(7,\underline{2},\underline{2},3),(7,\underline{2},\underline{2},5)\}},$$

$$\underbrace{\big(3, \{(3,3)\}, \{(3,8)\}\big)}_{d=3,\{(3,\underline{3},\underline{3},8)\}}, \underbrace{\big(6, \{(7,6)\}, \{\}\big)}_{d=6,\{\}}, \underbrace{\big(5, \{\}, \{(5,4)\}\big)}_{d=5,\{\}} \Big\}$$

IMDb (Teil 2)

und damit zurück zu:

2. Was sind die Datenmanagement und -analyseprobleme dahinter?

Frage 3

Wie stellen wir Anfragen an diese Daten?

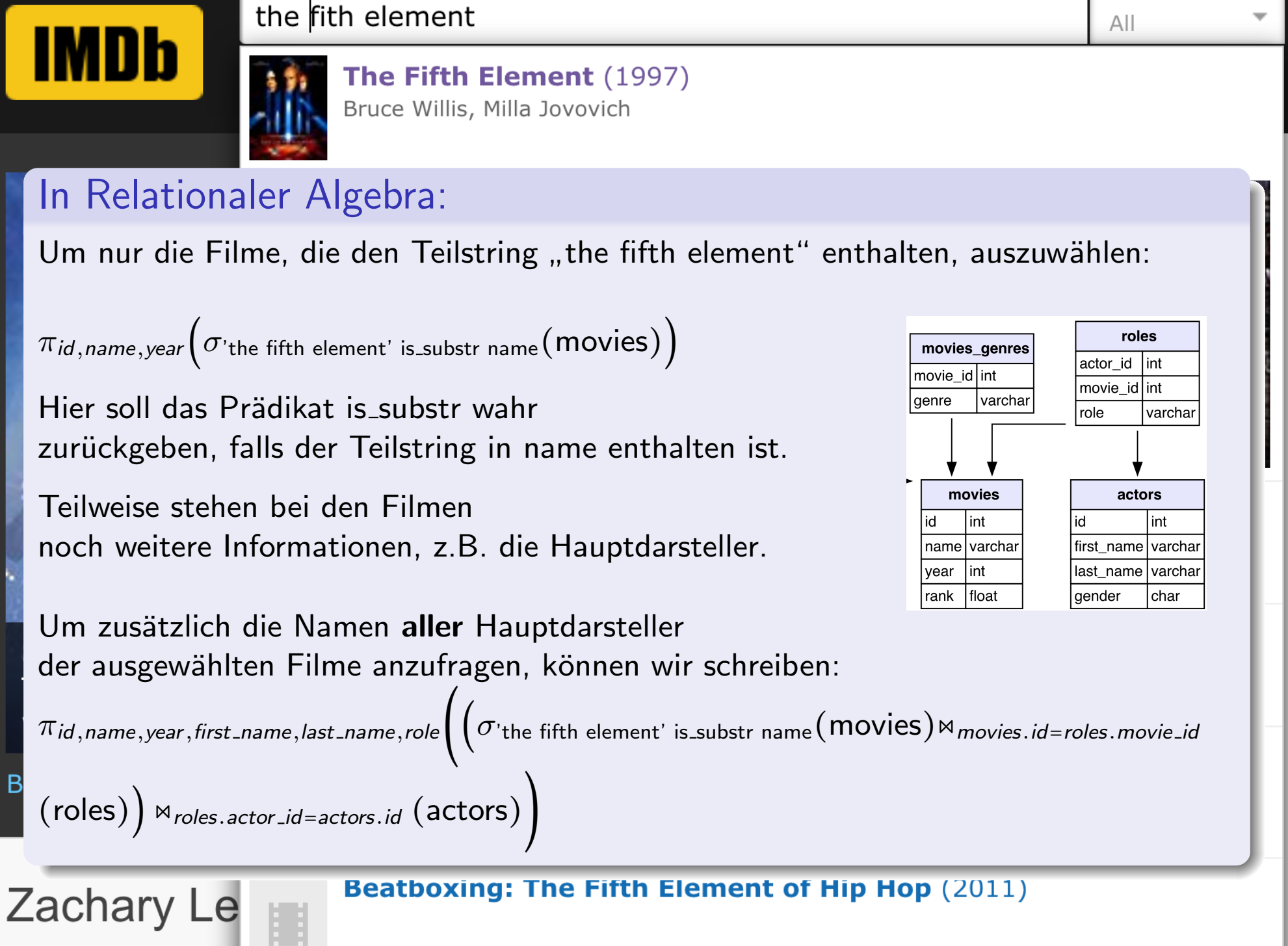
Mit der relationalen Algebra!

Transfer auf IMDb

4. Transfer der Grundlagen auf die konkrete Anwendung

Das haben wir diese und vergangene Woche für ER, das relationale Modell und die relationale Algebra bereits ‘inline’ gemacht in zahlreichen Beispielen.

Aber nochmal zurück zu unserem Anfangsbeispiel:



In Relationaler Algebra:

Um nur die Filme, die den Teilstring „the fifth element“ enthalten, auszuwählen:

$$\pi_{id,name,year}\left(\sigma_{\text{'the fifth element' is\_substr name}}(\text{movies})\right)$$

Hier soll das Prädikat is_substr wahr zurückgeben, falls der Teilstring in name enthalten ist.

Teilweise stehen bei den Filmen noch weitere Informationen, z.B. die Hauptdarsteller.

Um zusätzlich die Namen **aller** Hauptdarsteller der ausgewählten Filme anzufragen, können wir schreiben:

$$\pi_{id,name,year,first\_name,last\_name,role}\left(\left(\sigma_{\text{'the fifth element' is\_substr name}}(\text{movies}) \bowtie_{movies.id=roles.movie\_id} (\text{roles})\right) \bowtie_{roles.actor\_id=actors.id} (\text{actors})\right)$$

Cast

Edit

Cast overview, first billed only:

Bruce Willis ... Korben Dallas

In Relationaler Algebra:

Der Click im Webbrowser auf "The Fifth Element" wählt aus der oben erzeugten Liste den Film mit der id=112205 aus.

Damit können wir die Anfrage so formulieren:

$$\pi_{first\_name,last\_name,role}\left(\sigma_{movies.id=112205}(\text{roles}) \bowtie_{roles.actor\_id=actors.id} (\text{actors})\right)$$

Problem:

Was wir **nicht** ausdrücken konnten in relationaler Algebra:

1. nicht alle Hauptdarsteller sondern nur k-viele
2. nur die tollsten, wichtigsten Schauspieler (first billed only)
3. die Reihenfolge der tollsten, wichtigsten Schauspieler
4. ob überhaupt Schauspieler angezeigt werden

Lee Evans ... Fog

Ausblick auf nächste Woche

Leider sind Ausdrücke in relationaler Algebra manchmal etwas unübersichtlich. Deswegen ist es eine andere Option, die relationale Algebra unter einer anderen Sprache zu verstecken, d.h. eine andere Anfragesprache zu konzipieren und diese dann jeweils automatisch in die relationale Algebra zu übersetzen.

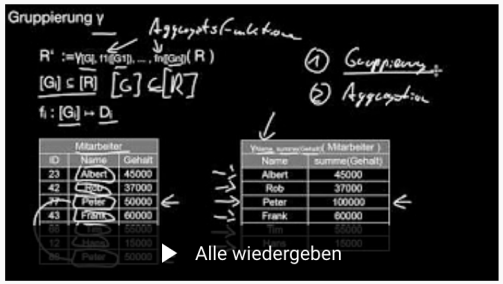
Fundamental Theorem of Software Engineering (FTSE):

„We can solve any problem by introducing an extra level of indirection . . . except for the problem of too many levels of indirection."

[David Wheeler]

Und genau das machen wir nächste Woche...

Weiterführendes Material



Relationale Algebra

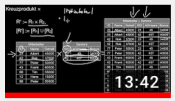
5 Videos • 33.225 Aufrufe • Zuletzt am 27.01.2014 aktualisiert

Öffentlich

Grundlagen der Relationalen Algebra

Prof. Dr. Jens Dittrich

SORTIEREN NACH

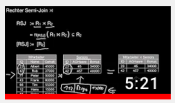


13.17 Relationale Algebra: Selektion, Projektion, Vereinigung, Differenz, Kreuzprodukt, Umbenennung

Prof. Dr. Jens Dittrich

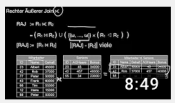
13.18a Relationale Algebra: Schnitt, Theta Join, Equi Join, Natural Join

Prof. Dr. Jens Dittrich



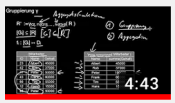
13.18b Relationale Algebra: Semi Joins, Anti Semi Joins

Prof. Dr. Jens Dittrich



13.18c Relationale Algebra: Äussere Joins

Prof. Dr. Jens Dittrich



13.18d Relationale Algebra: Gruppierung und Aggregation

Prof. Dr. Jens Dittrich

Youtube Videos von Prof. Dittrich zu Relationaler Algebra

sowie:

- Kapitel 3.4 in Kemper&Eickler (verfügbar in der Bibliothek)
- RelaX - relational algebra calculator
  - gist für IMDb_sample
  - gist für fotodb