

Umsetzung von Provenance-Anfragen in Big-Data-Analytics-Umgebungen

Tanja Auge

18.10.2017

Inhaltsverzeichnis

- 1 Problemstellung
- 2 Provenance
- 3 Das Konzept der CHASE-Inversen
- 4 Erste praktische Umsetzung
- 5 Zusammenfassung und Ausblick

Problemstellung

- Forschungsziel: Berechnung einer minimalen Teildatenbank, die die Ergebnisse der Auswertungs-Anfrage rekonstruieren kann, wobei
 - die Tupelanzahl der Originalrelation erhalten bleibt
 - und die Teildatenbank homomorph auf das Original abgebildet werden kann.
- Fragestellung: Was muss neben dem Ergebnis und der Anfrage noch archiviert werden?
 - Provenance-Polynome
 - Attributwerte der Quelltuplel
 - Ganze Tuplel der Quelldatenbank

Problemstellung

- Aufgabenstellung: Adaption der Techniken von Provenance-Anfragen *why*, *where*, *how* und *why not* in Umgebungen, die statt einfacher Anfragen wie Selektion, Projektion und Verbund auch OLAP-Operationen und weitere Machine-Learning-Algorithmen benutzen.
- Idee: Erweiterung bisheriger CHASE-Verfahren um eine BACKCHASE-Phase sowie den Aspekt der Provenance-Analyse

Fortlaufendes Beispiel

```
SELECT Matrikelnr, Modulnr, Note  
FROM Studenten JOIN Noten  
ON (Studenten.Matrikelnr = Noten.Matrikelnr)  
WHERE Studenten.Vorname = 'Max'
```

Matrikelnr	Modulnr	Note
3	2	2.3
7	2	3.3
3	4	1.3
7	5	1.7
3	7	1.7

Provenance-Anfragen und -Antworten

Anfrage-Typ	Fragestellung
<i>where</i>	Woher kommen die Daten?
<i>why</i>	Warum dieses Ergebnis?
<i>how</i>	Wie kommt das Ergebnis zustande?
<i>why not</i>	Warum fehlt ein bestimmtes Element im Ergebnis?

Antwort-Typ	Ergebnis
extensional	Tupel aus den Originaldaten
intensional	Beschreibung der Daten
anfragebasiert	Selektionsprädikate
modifikationsbasiert	Vorschlag zur minimalen Änderung der Auswertung

Provenance-Anfragen und -Antworten

Anfrage-Typ	Fragestellung
<i>where</i>	Woher kommen die Daten?
<i>why</i>	Warum dieses Ergebnis?
<i>how</i>	Wie kommt das Ergebnis zustande?
<i>why not</i>	Warum fehlt ein bestimmtes Element im Ergebnis?

Antwort-Typ	Ergebnis
<i>extensional</i>	Tupel aus den Originaldaten
intensional	Beschreibung der Daten
anfragebasiert	Selektionsprädikate
modifikationsbasiert	Vorschlag zur minimalen Änderung der Auswertung

Zeugenbasis (*why*-Provenance)

Definition

- Ein **Zeuge** w eines Ergebnistupels $t \in Q(d)$ ist eine Teilrelation $w \subseteq d$ mit $t \in Q(w)$.
- Eine Menge W von Zeugen w_i eines Ergebnistupels t heißt **Zeugenmenge**.
- Die **Zeugenbasis** entspricht der Menge aller Zeugenmengen.

Zeugenbasis (*why*-Provenance)

Definition

- Ein **Zeuge** w eines Ergebnistupels $t \in Q(d)$ ist eine Teilrelation $w \subseteq d$ mit $t \in Q(w)$.
- Eine Menge W von Zeugen w_i eines Ergebnistupels t heißt **Zeugenmenge**.
- Die **Zeugenbasis** entspricht der Menge aller Zeugenmengen.

Beispiel:

- Zeugenmenge des ersten Ergebnistupel: $W_1 = \{S_3, N_7\}$
- Zeugenbasis:
 $W = \{\{S_3, N_7\}, \{S_3, N_{13}\}, \{S_3, N_{20}\}, \{S_7, N_{11}\}, \{S_7, N_{16}\}\}$

Provenance-Polynome (*how*-Provenance)

- **Leere Relation:** $\emptyset(t) = 0$
- **Projektion:** $(\Pi_Y \rho)(t) := \sum_{t=t' \text{ auf } Y} \rho(t') \neq 0 \rho(t')$
- **Selektion:** $(\sigma_P \rho)(t) := \rho(t) \cdot_K \mathbf{P}(t)$
- **Vereinigung:** $(\rho_1 \cup_K \rho_2)(t) := \rho_1(t) +_K \rho_2(t)$
- **Natürlicher Verbund:** $(\rho_1 \bowtie \rho_2)(t) := \rho_1(t_1) \cdot_K \rho_2(t_2)$
- **Umbenennung:** $(\zeta_\beta \rho)(t) := \rho(t \circ \beta)$
- **Aggregationsergebnis:**

$$\rho'(t) = \begin{cases} \gamma_K(\sum_{t' \in T} \rho(t')), & \text{falls } T \neq \emptyset \text{ und} \\ & \forall x \in Z : t(x) = \sum_{t' \in T} \rho(t') \otimes t'(x) \\ 0, & \text{sonst} \end{cases}$$

Zusammenhang: *why*, *where* und *how*

- Ordnung nach ihrem Informationsgehalt:

$$where \preceq why \preceq how$$

- Rekonstruktion der *where*- und *why*-Provenance aus der *how*-Provenance:

<i>where</i> -Provenance (Tabellenname)	<i>where</i> -Provenance (Tupelname)	<i>why</i> -Provenance	<i>how</i> -Provenance
Studenten	S_3	(S_3, N_7)	$S_3 \cdot N_7$



Rekonstruktion der *where*- und *why*-Provenance aus der *how*-Provenance

- Provenance-Polynom (**how**-Provenance):

$$\frac{\overbrace{2.3 \otimes S_3 \cdot K N_7}^{p_1} +_{K \otimes M} \overbrace{1.3 \otimes S_3 \cdot K N_{13}}^{p_2} +_{K \otimes M} \overbrace{1.7 \otimes S_3 \cdot K N_{20}}^{p_3} +_{K \otimes M} \overbrace{3.3 \otimes S_7 \cdot K N_{11}}^{p_4} +_{K \otimes M} \overbrace{4.0 \otimes S_7 \cdot K N_{16}}^{p_5}}{\underbrace{S_3 \cdot K N_7}_{p_1} +_{K \otimes M} \underbrace{S_3 \cdot K N_{13}}_{p_2} +_{K \otimes M} \underbrace{S_3 \cdot K N_{20}}_{p_3} +_{K \otimes M} \underbrace{S_7 \cdot K N_{11}}_{p_4} +_{K \otimes M} \underbrace{S_7 \cdot K N_{16}}_{p_5}}$$

- Zeugenbasis (**why**-Provenance):

$$\left\{ \underbrace{\{S_3, N_7\}}_{p_1}, \underbrace{\{S_3, N_{13}\}}_{p_2}, \underbrace{\{S_3, N_{20}\}}_{p_3}, \underbrace{\{S_7, N_{11}\}}_{p_4}, \underbrace{\{S_7, N_{16}\}}_{p_5} \right\}$$

- where**-Provenance:

STUDENTEN, NOTEN

Der CHASE-Algorithmus

- Universalwerkzeug
- Arbeitet \star in \bigcirc ein, d.h.

$$\text{chase}_{\star}(\bigcirc) = \bigstar$$

- Diverse Varianten und Anwendungsmöglichkeiten:

	\star	\bigcirc	Ergebnis	Ziel
I.	Abhängigkeiten	Anfragen	Anfragen	Semantische Optimierung
II.	Sichten	Anfragen	Anfragen auf Sichten	AQuV
III.	s-t tg, egd	Quell-DB	Ziel-DB	Datenaustausch, Datenintegration

Idee für ein neues CHASE&BACKCHASE-Verfahren

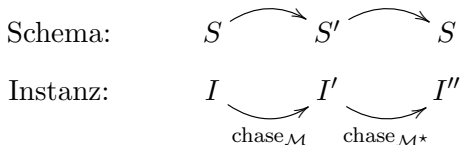
- Ziel: Erweiterung der dritten Variante auf beliebige Transformationen
- Vorgehen: Kombination der Techniken II. und III.
 - Grundlage: CHASE-Variante III.
 - Erweiterung: BACKCHASE-Phase aus II.
- Existenz einer CHASE-inversen Schemaabbildung nicht zwangsläufig gewährleistet
- Erweiterung der CHASE- und BACKCHASE-Phase um den Provenance-Aspekt

CHASE&BACKCHASE-Verfahrens für den Nachweis von CHASE-Inversen

Definition

Seien zwei Schemaabbildungen $\mathcal{M} = (S, S', \Sigma)$ und $\mathcal{M}^* = (S', S, \Sigma')$ gegeben. Dann gilt:

- CHASE: Berechne den CHASE von I bzgl. \mathcal{M} als Sequenz von s-t tg- und egd-Regeln.
- BACKCHASE: Berechne den CHASE von I' bzgl. \mathcal{M}^* als Sequenz von s-t tg- und egd-Regeln.



s-t tgd- und egd-Regel

Definition (s-t tgd-Regel)

Kombination von Quell tupeln s in Zieletupel t

Definition (egd-Regel)

Ersetze alle Nullwerte durch entsprechende

- Konstanten
- Nullwerte

GLAV-Abbildung

Definition (GLAV-Abbildung)

Eine **global-and-local-as-view**-Abbildung entspricht der Form:

$$\forall x : (\Phi(x) \rightarrow \exists y : \Psi(x, y))$$

GLAV-Abbildung

Definition (GLAV-Abbildung)

Eine **global-and-local-as-view**-Abbildung entspricht der Form:

$$\forall x : (\Phi(x) \rightarrow \exists y : \Psi(x, y))$$

Beispiel:

- Projektion: $\forall a_1, a_2, a_3 : (r(a_1, a_2, a_3) \rightarrow r(a_1, a_3))$

Exakte CHASE-Inverse

Definition

Sei $\mathcal{M} = (S, S', \Sigma)$ eine GLAV-Schemaabbildung. Dann heißt $\mathcal{M}^* = (S', S, \Sigma')$ **exakte CHASE-Inverse** für \mathcal{M} , wenn für jede Instanz I über S gilt:

$$I = \text{chase}_{\mathcal{M}^*}(\text{chase}_{\mathcal{M}}(I)).$$

Exakte CHASE-Inverse

Definition

Sei $\mathcal{M} = (S, S', \Sigma)$ eine GLAV-Schemaabbildung. Dann heißt $\mathcal{M}^* = (S', S, \Sigma')$ **exakte CHASE-Inverse** für \mathcal{M} , wenn für jede Instanz I über S gilt:

$$I = \text{chase}_{\mathcal{M}^*}(\text{chase}_{\mathcal{M}}(I)).$$

- Notwendige Bedingung: Übereinstimmung von Quell- und Urinstanz, d.h. $I = I''$
- Hinreichende Bedingung: —

CHASE-Inverse

Definition

Sei $\mathcal{M} = (S, S', \Sigma)$ eine GLAV-Schemaabbildung. Dann heißt $\mathcal{M}^* = (S', S, \Sigma')$ **CHASE-Inverse** für \mathcal{M} , wenn für jede Instanz I über S gilt:

$$I \leftrightarrow \text{chase}_{\mathcal{M}^*}(\text{chase}_{\mathcal{M}}(I)).$$

CHASE-Inverse

Definition

Sei $\mathcal{M} = (S, S', \Sigma)$ eine GLAV-Schemaabbildung. Dann heißt $\mathcal{M}^* = (S', S, \Sigma')$ **CHASE-Inverse** für \mathcal{M} , wenn für jede Instanz I über S gilt:

$$I \leftrightarrow \text{chase}_{\mathcal{M}^*}(\text{chase}_{\mathcal{M}}(I)).$$

- Notwendige Bedingung: Äquivalenz von Quell- und Urinstanz, d.h. $I \leftrightarrow I''$
- Hinreichende Bedingung: Existenz einer exakten CHASE-Inversen

Ergebnisäquivalenz, Ausschnitt

Definition

Sei $\mathcal{M} = (S, S', \Sigma)$ eine GLAV-Schemaabbildung. Seien I und J zwei Instanzen über S . Dann sind I und J **ergebnisäquivalent** bzgl. \mathcal{M} , wenn gilt:

$$\text{chase}_{\mathcal{M}}(I) \leftrightarrow \text{chase}_{\mathcal{M}}(J).$$

Schreibweise: $I \leftrightarrow_{\mathcal{M}} J$

Ergebnisäquivalenz, Ausschnitt

Definition

Sei $\mathcal{M} = (S, S', \Sigma)$ eine GLAV-Schemaabbildung. Seien I und J zwei Instanzen über S . Dann sind I und J **ergebnisäquivalent** bzgl. \mathcal{M} , wenn gilt:

$$\text{chase}_{\mathcal{M}}(I) \leftrightarrow \text{chase}_{\mathcal{M}}(J).$$

Schreibweise: $I \leftrightarrow_{\mathcal{M}} J$

Definition

Sei I Quellinstanz mit Tupeln $(a_{i_1}, \dots, a_{i_n})$. Eine Instanz I'' heißt **Ausschnitt** von I , wenn ihre Tupel der Form $(x_{i_1}, \dots, x_{i_n})$ mit $x_{i_j} = a_{i_j}$ oder $x_{i_j} = n_j$ sind.

Schreibweise: $I'' \preceq I$

Relaxte CHASE-Inverse

Definition

Sei $\mathcal{M} = (S, S', \Sigma)$ eine GLAV-Schemaabbildung. Dann heißt \mathcal{M}^* **relaxte CHASE-Inverse** von \mathcal{M} , wenn für jede Instanz I über S gilt:

- $I'' \leftrightarrow_{\mathcal{M}} I$
- $I'' \rightarrow I$

Dabei ist $I'' = \text{chase}_{\mathcal{M}^*}(\text{chase}_{\mathcal{M}}(I))$.

Relaxte CHASE-Inverse

Definition

Sei $\mathcal{M} = (S, S', \Sigma)$ eine GLAV-Schemaabbildung. Dann heißt \mathcal{M}^* **relaxte CHASE-Inverse** von \mathcal{M} , wenn für jede Instanz I über S gilt:

- $I'' \leftrightarrow_{\mathcal{M}} I$
- $I'' \rightarrow I$

Dabei ist $I'' = \text{chase}_{\mathcal{M}^*}(\text{chase}_{\mathcal{M}}(I))$.

- Notwendige Bedingung: Erhalt der Tupelmenge der Urinstanz, d.h. $I'' \preceq I$
- Hinreichende Bedingung: Existenz einer CHASE-Inversen

Ergebnisäquivalente CHASE-Inverse

Definition

Seien $\mathcal{M} = (S, S', \Sigma)$ eine GLAV-Schemaabbildung von S nach S' . Dann heißt \mathcal{M}^* **ergebnisäquivalente CHASE-Inverse** von \mathcal{M} , wenn für jede Instanz I und $I'' = \text{chase}_{\mathcal{M}^*}(\text{chase}_{\mathcal{M}}(I))$ gilt:

$$I'' \leftrightarrow_{\mathcal{M}} I.$$

Ergebnisäquivalente CHASE-Inverse

Definition

Seien $\mathcal{M} = (S, S', \Sigma)$ eine GLAV-Schemaabbildung von S nach S' . Dann heißt \mathcal{M}^* **ergebnisäquivalente CHASE-Inverse** von \mathcal{M} , wenn für jede Instanz I und $I'' = \text{chase}_{\mathcal{M}^*}(\text{chase}_{\mathcal{M}}(I))$ gilt:

$$I'' \leftrightarrow_{\mathcal{M}} I.$$

- Notwendige Bedingung: —
- Hinreichende Bedingung: Existenz einer relaxten CHASE-Inversen.

CHASE-Inverse Schemaabbildungen

- Ordnung der CHASE-Inversen nach ihrem Informationsgehalt

ergebnisäquivalent \preceq relaxt
 \preceq CHASE-Inverse
 \preceq exakt

- Angabe von CHASE-Inversen für:
 - Algebraische Grundoperationen
 - Kompositionen von Grundoperationen
 - OLAP-Operationen

Algebraische Grundoperationen mit CHASE-Inversen

Operation:	Inverse Abbildung ohne Provenance-Information:	Inverse Abbildung mit Provenance-Information:
$r(\mathcal{R})$	Exakt	Exakt
$\beta_{A_j \leftarrow A_i}(r(\mathcal{R}))$	Exakt	Exakt
$\pi_{A_i}(r(\mathcal{R}))$	Relaxt	Relaxt
	Ergebnisäquivalent	Relaxt
$r_1(\mathcal{R}_1) \bowtie r_2(\mathcal{R}_2)$	Exakt	Exakt
	Ergebnisäquivalent	Ergebnisäquivalent
$\sigma_{A_i \theta c}(r(\mathcal{R}))$ mit $\theta \in \{<, \leq, =, \geq, >\}$	Ergebnisäquivalent	Ergebnisäquivalent
$\sigma_{A_i \theta A_j}(r(\mathcal{R}))$ mit $\theta \in \{<, \leq, =, \geq, >\}$	Ergebnisäquivalent	Ergebnisäquivalent
$\sigma_{A_i \neq c}(r(\mathcal{R}))$	xxx	xxx
$\sigma_{A_i \neq A_j}(r(\mathcal{R}))$	xxx	xxx
$r_1(\mathcal{R}_1) \cup r_2(\mathcal{R}_2)$	Ergebnisäquivalent	Exakt
$r_1(\mathcal{R}_1) \cap r_2(\mathcal{R}_2)$	Ergebnisäquivalent	Ergebnisäquivalent
$r_1(\mathcal{R}_1) - r_2(\mathcal{R}_2)$	xxx	xxx
MAX ($r(\mathcal{R})$) / MIN ($r(\mathcal{R})$)	Ergebnisäquivalent	Ergebnisäquivalent
COUNT ($r(\mathcal{R})$)	Relaxt	Relaxt
SUM ($r(\mathcal{R})$)	xxx	Exakt
AVG ($r(\mathcal{R})$)	xxx	Exakt
$\gamma_{G; F_j(A_j)}(r(\mathcal{R}))$	Ergebnisäquivalent	Ergebnisäquivalent
	Relaxt	Relaxt
	xxx	Exakt
$r(\mathcal{R}) \theta \alpha$ mit $\theta \in \{+, -, \cdot, : \}$	Exakt	Exakt
$r(\mathcal{R}) \bmod \alpha$	Ergebnisäquivalent	Ergebnisäquivalent

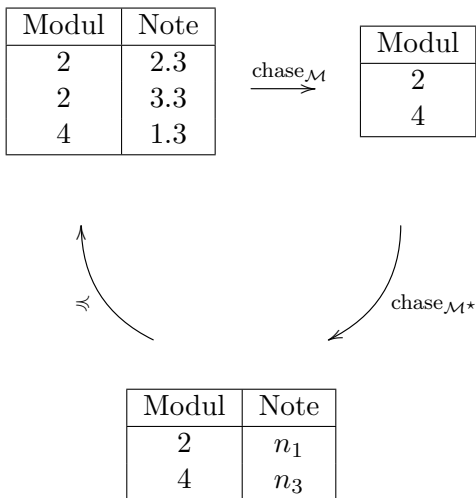
Algebraische Grundoperationen mit CHASE-Inversen

Operation:	Inverse Abbildung ohne Provenance-Information:	Inverse Abbildung mit Provenance-Information:
$r(\mathcal{R})$	Exakt	Exakt
$\beta_{A_i \leftarrow A_j}(r(\mathcal{R}))$	Exakt	Exakt
$\pi_{A_i}(r(\mathcal{R}))$	Relaxt	Relaxt
	Ergebnisäquivalent	Relaxt
$r_1(\mathcal{R}_1) \bowtie r_2(\mathcal{R}_2)$	Exakt	Exakt
	Ergebnisäquivalent	Ergebnisäquivalent
$\sigma_{A_i \theta c}(r(\mathcal{R}))$ mit $\theta \in \{<, \leq, =, \geq, >\}$	Ergebnisäquivalent	Ergebnisäquivalent
$\sigma_{A_i \theta A_j}(r(\mathcal{R}))$ mit $\theta \in \{<, \leq, =, \geq, >\}$	Ergebnisäquivalent	Ergebnisäquivalent
$\sigma_{A_i \neq c}(r(\mathcal{R}))$	xxx	xxx
$\sigma_{A_i \neq A_j}(r(\mathcal{R}))$	xxx	xxx
$r_1(\mathcal{R}_1) \cup r_2(\mathcal{R}_2)$	Ergebnisäquivalent	Exakt
$r_1(\mathcal{R}_1) \cap r_2(\mathcal{R}_2)$	Ergebnisäquivalent	Ergebnisäquivalent
$r_1(\mathcal{R}_1) - r_2(\mathcal{R}_2)$	xxx	xxx
MAX ($r(\mathcal{R})$) / MIN ($r(\mathcal{R})$)	Ergebnisäquivalent	Ergebnisäquivalent
COUNT ($r(\mathcal{R})$)	Relaxt	Relaxt
SUM ($r(\mathcal{R})$)	xxx	Exakt
AVG ($r(\mathcal{R})$)	xxx	Exakt
$\gamma_{G_i; F_j(A_j)}(r(\mathcal{R}))$	Ergebnisäquivalent	Ergebnisäquivalent
	Relaxt	Relaxt
	xxx	Exakt
$r(\mathcal{R}) \theta \alpha$ mit $\theta \in \{+, -, \cdot, : \}$	Exakt	Exakt
$r(\mathcal{R}) \bmod \alpha$	Ergebnisäquivalent	Ergebnisäquivalent

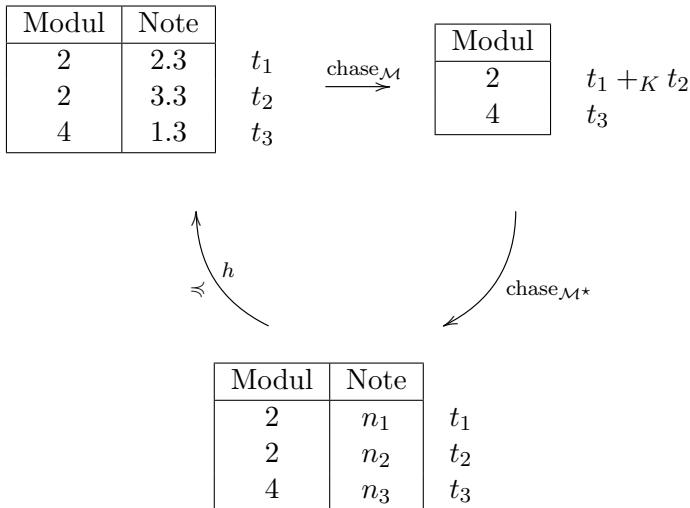
Projektion

- Probleme / Schwierigkeiten: Duplikate
- Ergebnisäquivalente CHASE-Inverse: Ohne Provenance-Informationen
- Relaxte CHASE-Inverse: Mit Provenance-Informationen
- CHASE-Inverse: Kann nicht explizit angegeben werden
- Exakte CHASE-Inverse: Merken „wegprojizierter“ Attribute

Projektion



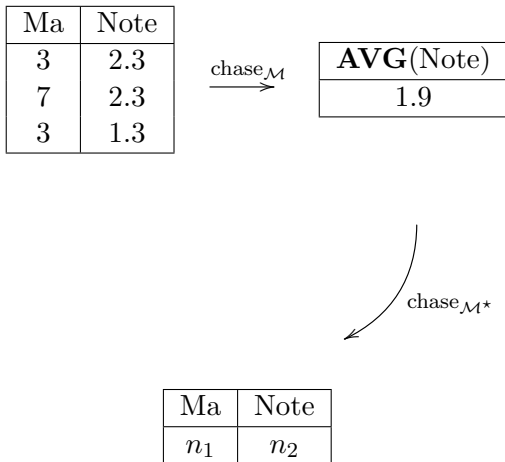
Projektion



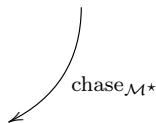
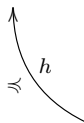
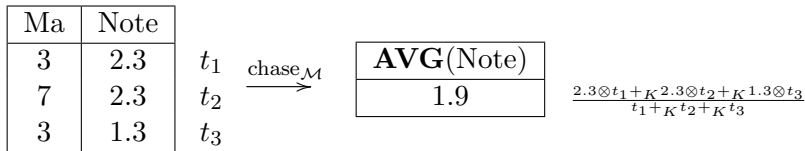
Aggregatfunktion AVG

- Probleme / Schwierigkeiten: Keine
- Ergebnisäquivalente CHASE-Inverse: Kann nicht explizit angegeben werden
- Relaxte CHASE-Inverse: Kann nicht explizit angegeben werden
- CHASE-Inverse: Kann nicht explizit angegeben werden
- Exakte CHASE-Inverse: Mit Provenance-Informationen

Aggregatfunktion AVG



Aggregatfunktion AVG



Ma	Note
n_1	2.3
n_2	2.3
n_3	1.3

t_1
 t_2
 t_3

Erste praktische Umsetzung

- Erweiterungen des Programms **ProSA**
- **ProSA**: Prototyp des Projekts *NEidI*¹

```
Inverse.txt
Inversen-Typ:
Ergebnisäquivalente CHASE-Inverse

Zeugenmenge:
[N11, N13, N16, N20, N7, S3, S3, S3, S7, S7]

Zeugenliste:
[N11, N13, N16, N20, N7, S3, S7]

Quelltupel der Zeugenliste:
studenten-Relation:
matrikelnr; name; vorname; studiengang; id_studenten
3; Müller; Max; Elektrotechnik; S3
7; Mustermann; Max; Elektrotechnik; S7
-----
noten-Relation:
modulnr; matrikelnr; semester; note; id_noten
2; 7; WS 15/16; 3.3; N11
4; 3; WS 16/17; 1.3; N13
5; 7; SS 17; 1.7; N16
7; 3; SS 16; 1.7; N20
2; 3; WS 14/15; 2.3; N7
```

¹Sabrina Brossmann, Pia Wilsdorf, Tanja Auge, 2016

Zusammenfassung der Arbeit

- Zusammenhang zwischen *why*-, *where*- und *how*-Provenance
- Definition der ergebnisäquivalenten CHASE-Inversen
- Aufstellen eines CHASE&BACKCHASE-Verfahrens für den Nachweis von CHASE-Inversen
- Untersuchung der algebraischen Grundoperationen auf die Existenz von (exakten / relaxten / ergebnisäquivalenten) CHASE-Inversen
- Aufstellen von zusätzlichen Kriterien für die Existenz einer exakten CHASE-Inversen
- Erste praktische Umsetzung

Ausblick / Offene Probleme

- Definition der Provenance-Polynome für Differenzbildung sowie Selektion auf \neq
- Beschreibung der Aggregation und Gruppierung als s-t tgds
- Erweiterung des CHASE-Algorithmus für nicht-positive relationale Algebren
- Untersuchung intensionaler, anfragebasierter und modifikationsbasierter Antworten
- Automatische Berechnung der minimalen Zeugenbasis
- Erweiterung des Programms **ProSA** um Gruppierung, Vereinigung, ...

Vielen Dank für Ihre Aufmerksamkeit !