

Die Graham-Reduktion als Alternative zur Berechnung von Verbundpfaden bei der Anfragerekonstruktion

Masterarbeit
Hauke Schur



Gliederung

1. Motivation und Paleo
2. Graham-Reduktion und andere klassische Ansätze
3. Vergleich
4. Konzept der Arbeit
5. Fazit



1. Motivation

- Reverse Engineering von Datenbank Anfragen
- Paleo Framework für Top-K-Anfragen
- Ergebnisliste und originale Daten (oder einen Teil davon)
- neuer Ansatz für Verbundpfaderstellung

1. Paleo SQL Template

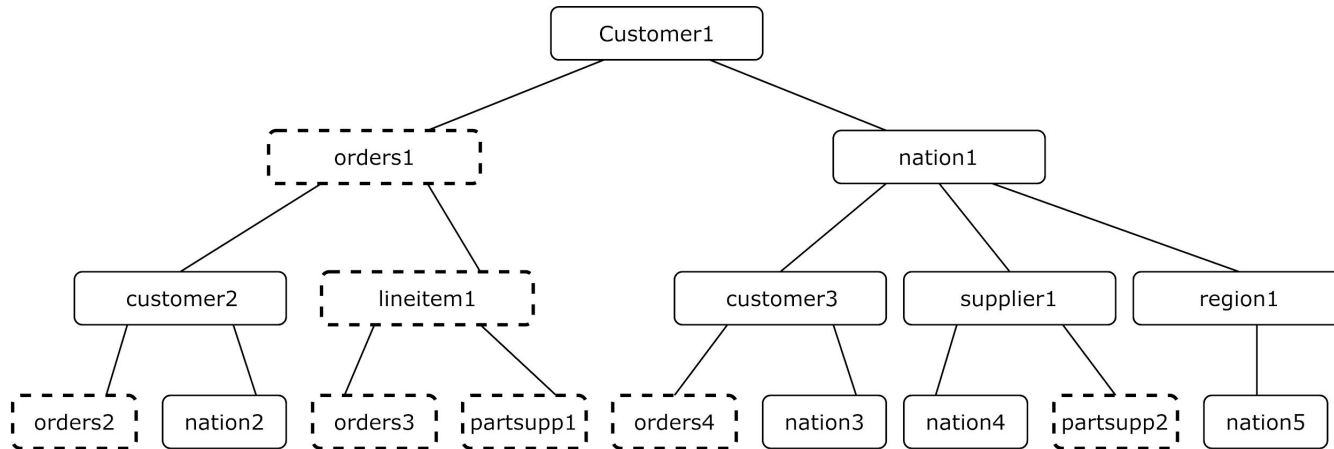
```
SELECT L.e, agg(value)
FROM R1, R2, ...
WHERE P1 and P2...
GROUP BY L.e
ORDER BY agg(value) DESC LIMIT k
```



1. Paleo-J

1. Untersuchen des Datenbankschemas
→ Identitätsattribut und Menge von Rankingattributen
2. Erstellung eines Suchbaums mit Schlüsselbeziehungen
- 3./4. Erstellung von Verbundpfaden
→ Verarbeitung einzelner Zweige des Suchbaums
5. Erzeugen von Anfragen
6. Überprüfen der Anfragen

1. Paleo-J Suchbaum



2. klassische Ansätze der Verbundpfaderstellung

- universal relation scheme assumption
- zugänglicher durch automatische Verbundpfaderstellung

Alternativen:

- Tableau: Darstellung in Tabellen, Verbundpfade über Homomorphismen/Abbildung
- Join Tree: direkte Ableitung der Verbundpfade aus einem Baum



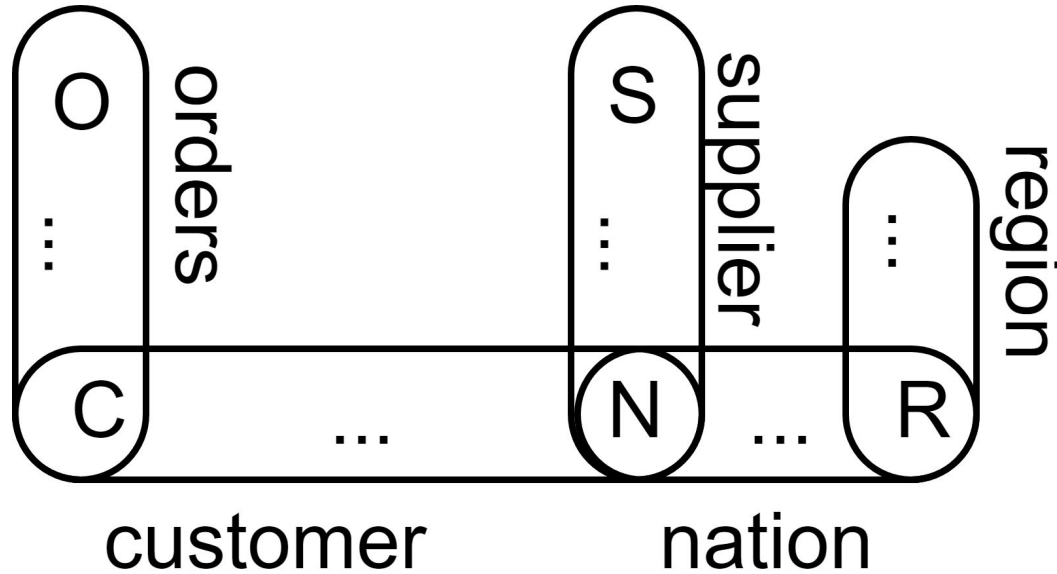
2. Graham-Reduktion von Hypergraphen

- gedacht um Hypergraph auf Azyklizität zu überprüfen
→ bei Erfolg gelten Eigenschaften wie Join Tree, Full Reducer, leichte 4NF
- mit Modifikation zur Verbundpfaderzeugung geeignet

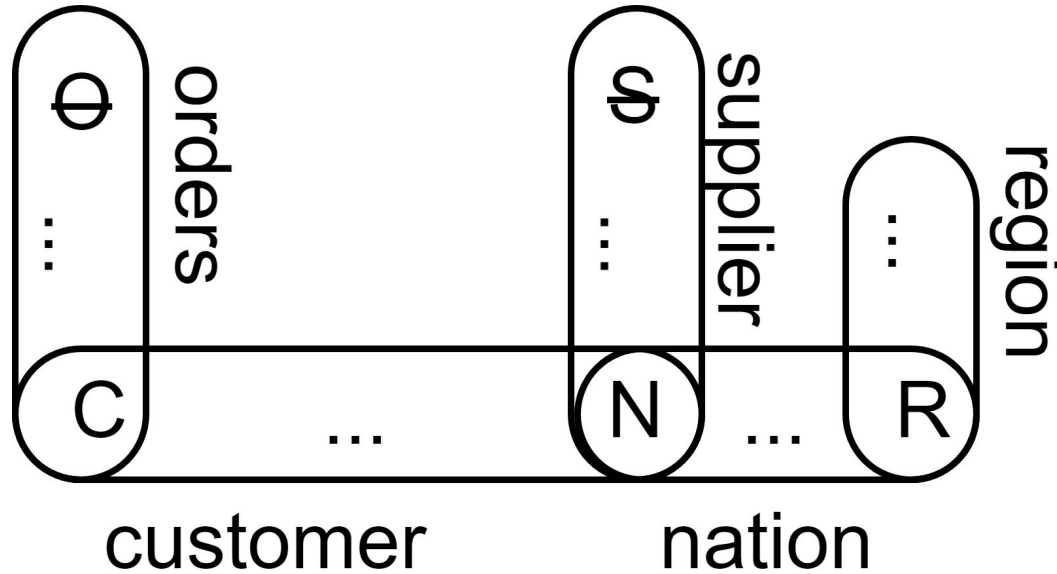
Graham-Reduktion:

1. Entfernen aller Knoten mit nur einer Kante
2. Entfernen aller Kanten, welche Teil einer anderen sind

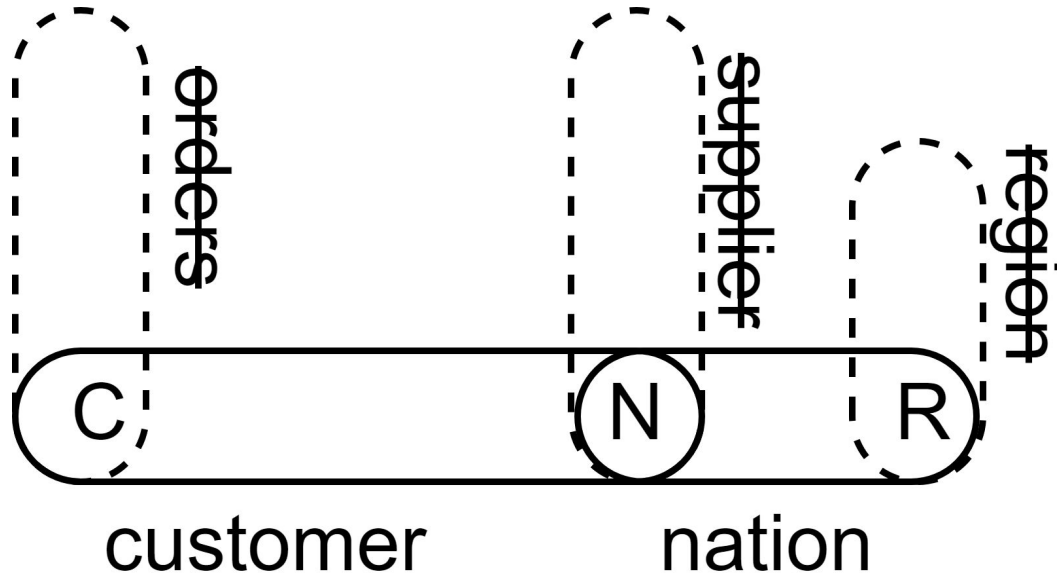
2. Graham-Reduktion



2. Graham-Reduktion



2. Graham-Reduktion



2. Graham-Reduktion



customer

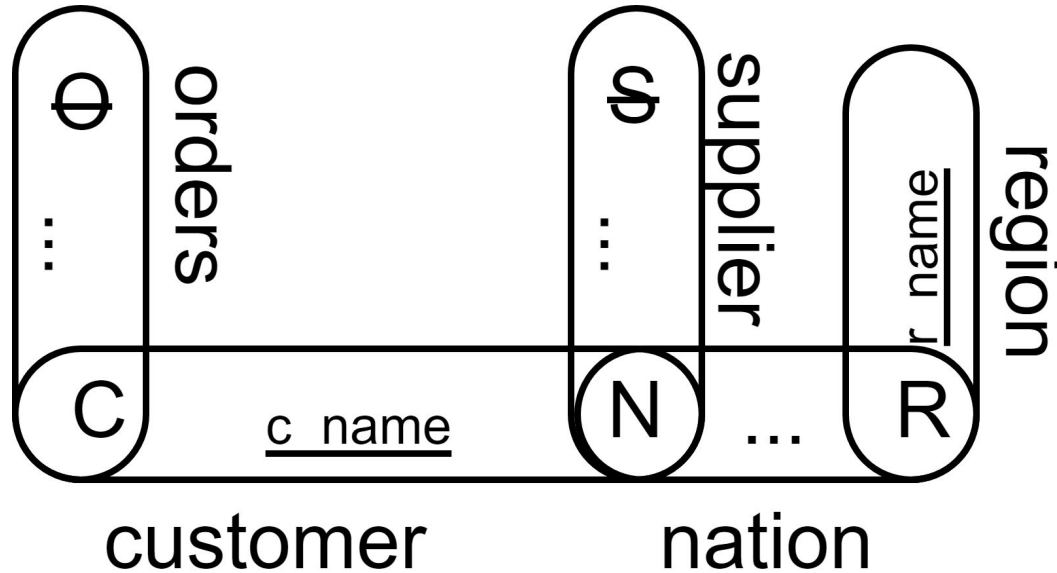
nation



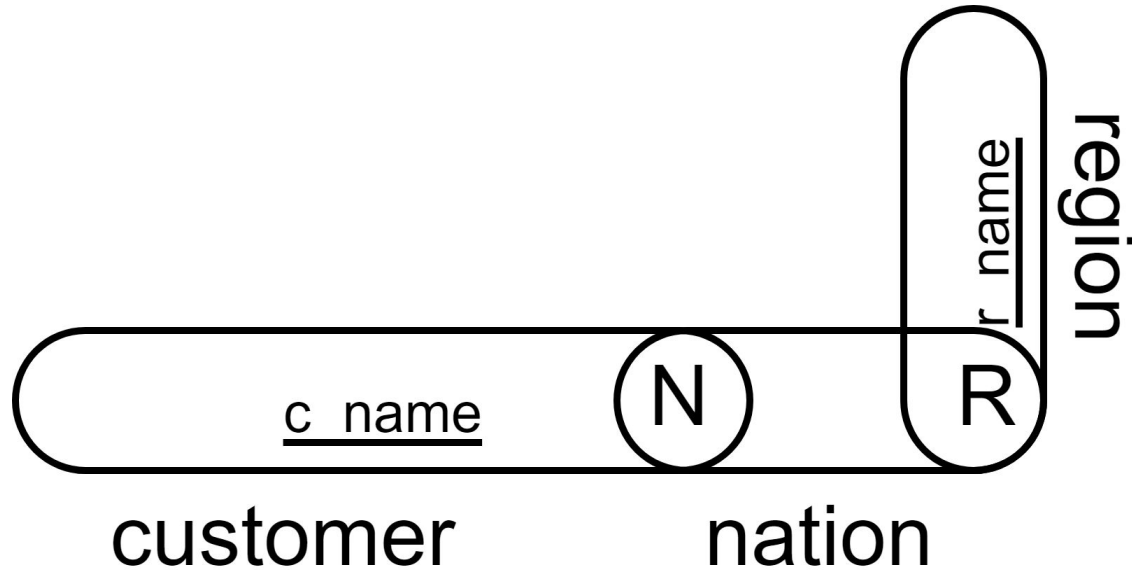
customer

nation

2. Graham-Reduktion mit sacred nodes



2. Graham-Reduktion mit sacred nodes



3. Vergleich Graham-Reduktion mit Paleo

- klassische Ansätze: gegebene Attribute \rightarrow ein Verbundpfad
- Paleo-J: Menge von möglichen Attributen \rightarrow alle Verbundpfade
- TPC-H als Referenzschema

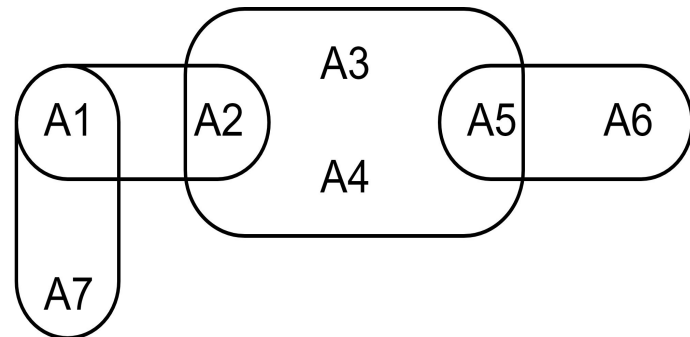
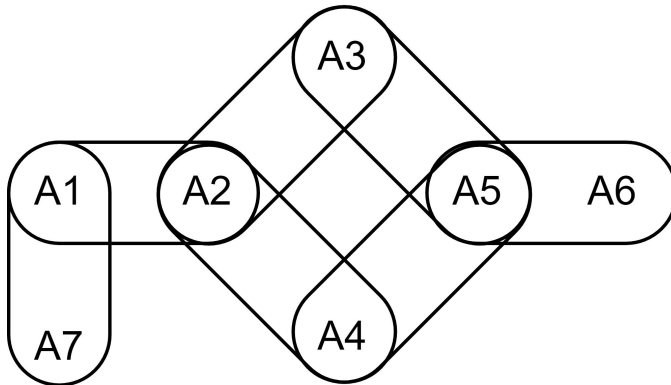
Probleme der Graham-Reduktion:

- zyklische Datenbankschemata
- Finden aller Verbundpfade nicht gesichert
- nur direkte Verbundpfade
- alle Relationen nur einfach

4. Konzept - Superkanten

Superkanten:

- Entfernen der Zyklen und Einfügen einer Superkante



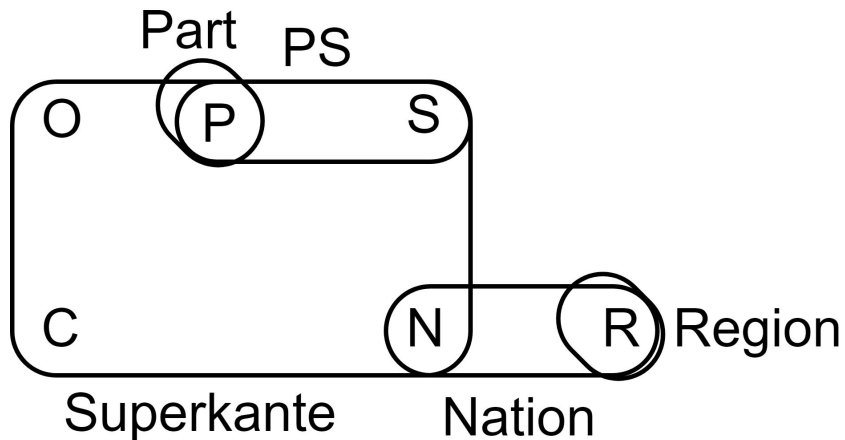
4. Konzept

Eliminierungshypergraph:

- Auffinden von Verbundpfaden mit verschluckten Relationen

Pfaderweiterung:

- Erweiterung plausibler Pfade um Nachbarn im Hypergraphen





4. Prototyp

- Kanten und Knoten enthalten sich je als Set

Ablauf

- Ermitteln von Zyklen → Auflösen als Superkanten mittels Dijkstra
- Erstellen der Verbundpfade für gegebene Attributmenge
- Einsatz des Eliminierungshypergraph
- Auflösen der Superkanten

4. Vergleich

eigener Ansatz:

- self joins nicht berücksichtigt / unvollständig
- abhängig von der Schemagröße

Paleo-J:

- vollständig
- abhängig von der Verbundpfadlänge



4. Tests

- Tests mit eigenen und TPC-H-Anfragen
- Paleo im Vorteil

Verbundpade der TPC-H-Anfragen:

- getestete Paleo-Version: 17 von 22 Pfaden
- eigener Ansatz: 8 (16) von 22 Pfaden



5. Fazit

- Vergleich schwer ohne gesamtes Framework
- Ansätze haben unterschiedliche Vor- und Nachteile
- eigener Ansatz möglicherweise schneller, aber unvollständig
- nur bedingt geeignet für Anfragenrekonstruktion



Vielen Dank für Ihre Aufmerksamkeit!