Universität Rostock

Traditio et Innovatio

Automatische Generierung von Domänengeneralisierungshierarchien

Bachelorarbeit - Endverteidigung

Richard Dabels

Motivation

Privatsphäre

- Unternehmen / Einrichtungen sammeln Informationen
- Betroffene erwarten Transparenz über Verwendung
- Gesetze fordern Datenschutz und -sparsamkeit

Motivation

Deanonymisierung / Gefahren

- Identifikation von Wählern in Cambridge Wahlliste
- Unternehmen, die Entscheidungen über Angestellte anhand medizinischer Unterlagen treffen
- Banken, die Kredite von Krebspatienten zurückfordern

Quelle: Sweeney, Latanya; *Guaranteeing anonymity when sharing medical data, the Datafly System*

Stand der Forschung

- Datafly
- k-Anonymität
- l-Diversity
- t-Closeness

Stand der Forschung

Datafly

- Eintrag anonym, wenn nicht von *k* anderen unterscheidbar
- Generalisierung numerischer Werte
- Löschung kategorischer Werte

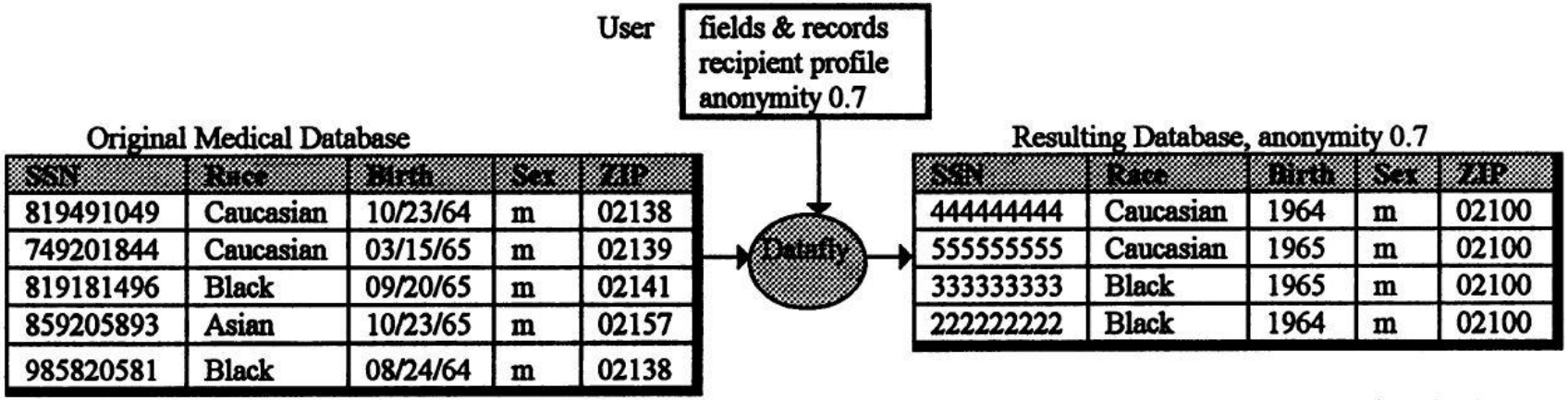


Abbildung: Sweeney, Latanya; *Guaranteeing anonymity when sharing medical data, the Datafly System*

Stand der Forschung

k-Anonymität

„Ein Datensatz ist von k-1 anderen nicht mehr unterscheidbar“

Studiengang	Geburtsdatum	PLZ	Ort	Note
Informatik	1989 - 1992	180xx	Rostock	3,2
Chemie	1989 - 1992	182xx	Bad Doberan	1,2
Informatik	1993 - 1996	180xx	Rostock	2,3
Informatik	1993 - 1996	180xx	Rostock	1,9
Chemie	1989 - 1992	182xx	Bad Doberan	2,5
Informatik	1989 - 1992	180xx	Rostock	2,3
Informatik	1989 - 1992	180xx	Rostock	1,9
Physik	1993 - 1996	181xx	Rostock	3,0
Informatik	1993 - 1996	180xx	Rostock	3,2
Physik	1993 - 1996	181xx	Rostock	2,9
Chemie	1989 - 1992	182xx	Bad Doberan	3,7
Physik	1993 - 1996	181xx	Rostock	3,0

Stand der Forschung

l-Diversity

k-Anonymität + *„mindestens l Vorkommen jedes sensitiven Wertes für alle Gruppen“*

Studiengang	Geburtsdatum	PLZ	Ort	Note
Informatik	1989 - 1992	180xx	Rostock	3,2
Informatik	1993 - 1996	180xx	Rostock	2,3
Informatik	1993 - 1996	180xx	Rostock	1,9
Informatik	1989 - 1992	180xx	Rostock	2,3
Informatik	1989 - 1992	180xx	Rostock	1,9
Informatik	1993 - 1996	180xx	Rostock	3,2

Stand der Forschung

t-Closeness

l-Diversity + *„Verteilung sensitiver Werte einer Gruppe ähnelt der Gesamtverteilung“*

Studiengang	Geburtsdatum	PLZ	Ort	Note
Informatik	1989 - 1992	180xx	Rostock	3,2
Chemie	1989 - 1992	182xx	Bad Doberan	1,2
Informatik	1993 - 1996	180xx	Rostock	2,3
Informatik	1993 - 1996	180xx	Rostock	1,9
Chemie	1989 - 1992	182xx	Bad Doberan	2,5
Informatik	1989 - 1992	180xx	Rostock	2,3
Informatik	1989 - 1992	180xx	Rostock	1,9
Physik	1993 - 1996	181xx	Rostock	3,0
Informatik	1993 - 1996	180xx	Rostock	3,2
Physik	1993 - 1996	181xx	Rostock	2,9
Chemie	1989 - 1992	182xx	Bad Doberan	3,7
Physik	1993 - 1996	181xx	Rostock	3,0

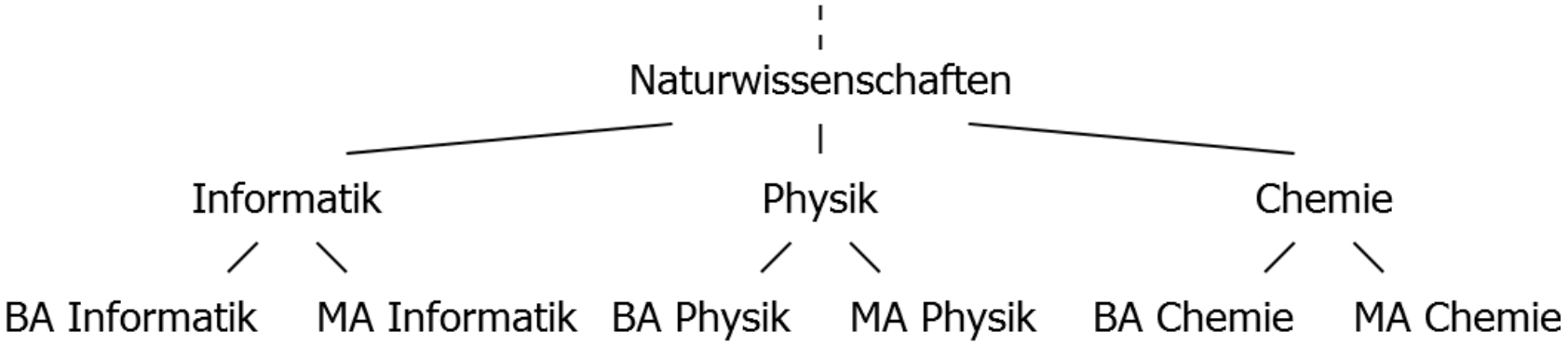
Stand der Technik

- Domänengeneralisierungshierarchien
- Clustering
- Ontologien

Stand der Technik

Domänengeneralisierungshierarchien

- Baumstrukturen zur Generalisierung von kategorischen Attributen



Stand der Technik

Domänengeneralisierungshierarchien

Matrikelnummer	Studiengang	Geburtsdatum	PLZ	Ort	Note
209202223	MA Informatik	22.06.1990	18057	Rostock	3,2
209204742	MA Chemie	14.04.1989	18209	Bad Doberan	1,7
212203190	BA Informatik	28.10.1995	18069	Rostock	1,9
212205002	BA Informatik	26.03.1993	18057	Rostock	3,7
212205999	MA Chemie	08.06.1991	18202	Bad Doberan	2,0
214200123	MA Informatik	01.01.1990	18057	Rostock	2,3
214201234	MA Informatik	06.07.1992	18059	Rostock	1,9
214202356	BA Physik	07.10.1995	18107	Rostock	3,0
214202644	MA Informatik	01.02.1993	18069	Rostock	3,2
214203141	BA Physik	17.07.1993	18147	Rostock	2,9
215202718	BA Chemie	22.02.1992	18209	Bad Doberan	2,5
215209159	MA Physik	25.12.1993	18119	Rostock	3,0

Stand der Technik

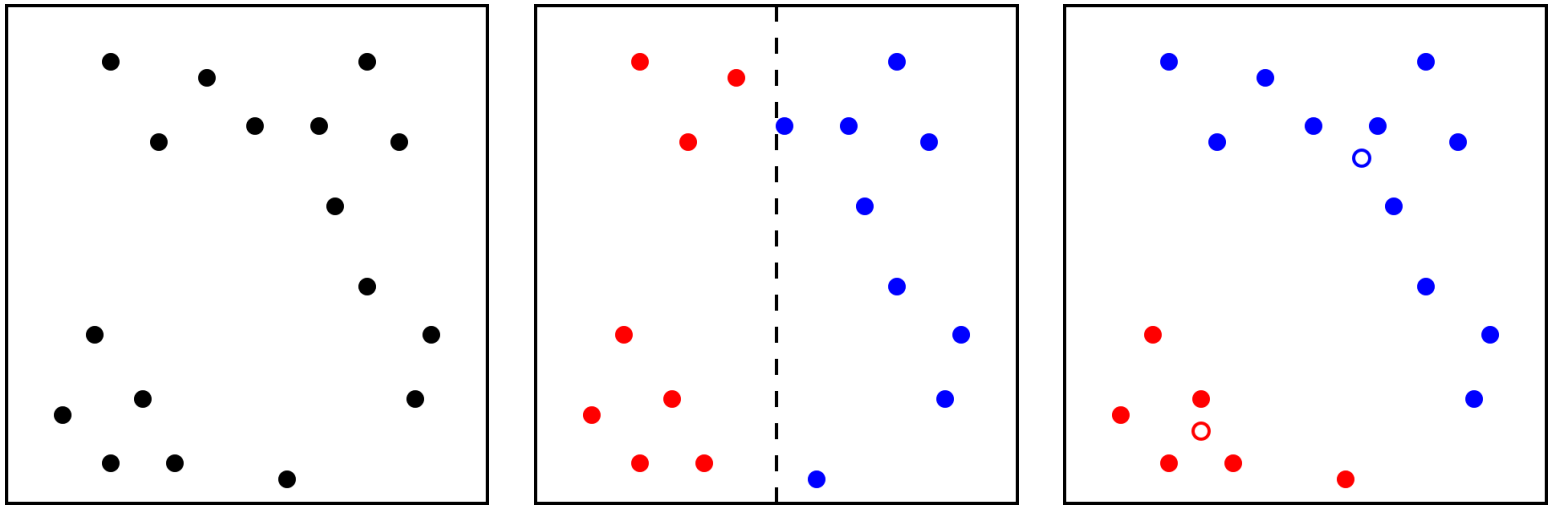
Domänengeneralisierungshierarchien

Studiengang	Geburtsdatum	PLZ	Ort	Note
Informatik	1989 - 1992	180xx	Rostock	3,2
Chemie	1989 - 1992	182xx	Bad Doberan	1,2
Informatik	1993 - 1996	180xx	Rostock	2,3
Informatik	1993 - 1996	180xx	Rostock	1,9
Chemie	1989 - 1992	182xx	Bad Doberan	2,5
Informatik	1989 - 1992	180xx	Rostock	2,3
Informatik	1989 - 1992	180xx	Rostock	1,9
Physik	1993 - 1996	181xx	Rostock	3,0
Informatik	1993 - 1996	180xx	Rostock	3,2
Physik	1993 - 1996	181xx	Rostock	2,9
Chemie	1989 - 1992	182xx	Bad Doberan	3,7
Physik	1993 - 1996	181xx	Rostock	3,0

Stand der Technik

Clustering

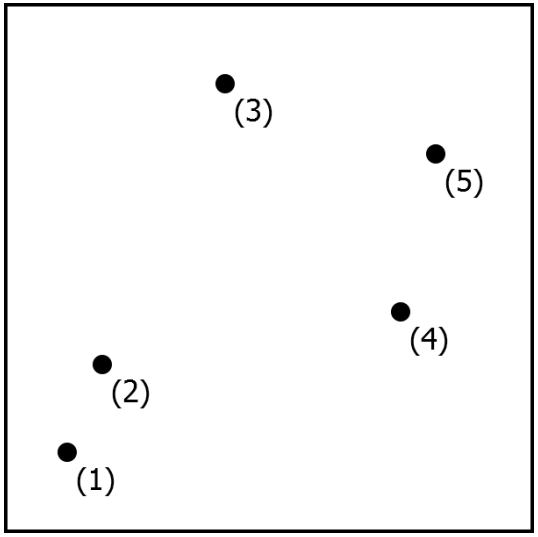
- k-Means-Clustering

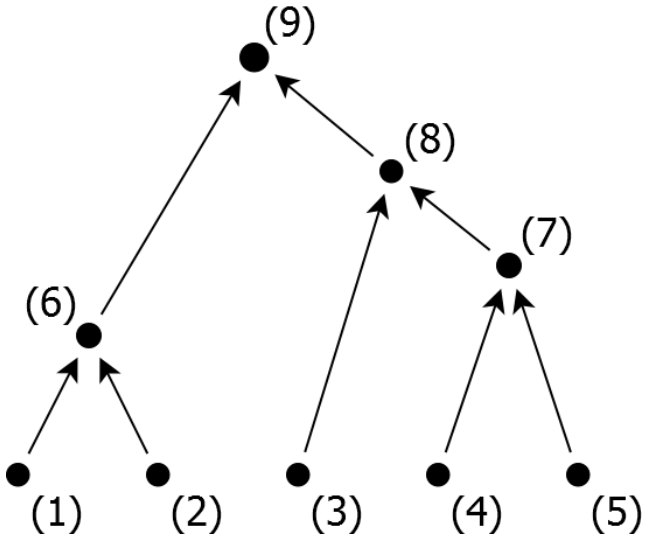


Stand der Technik

Clustering

- hierarchisches Clustering

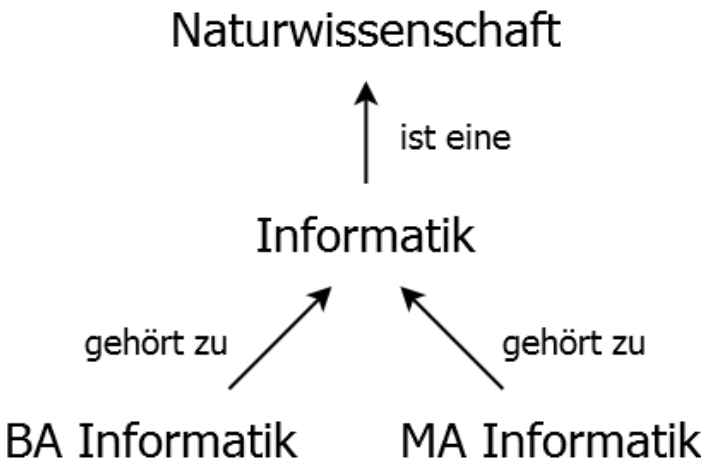




Stand der Technik

Ontologien

- basieren auf Beschreibungslogiken
- modellieren Semantik von Objekten sowie deren Beziehungen



Konzept

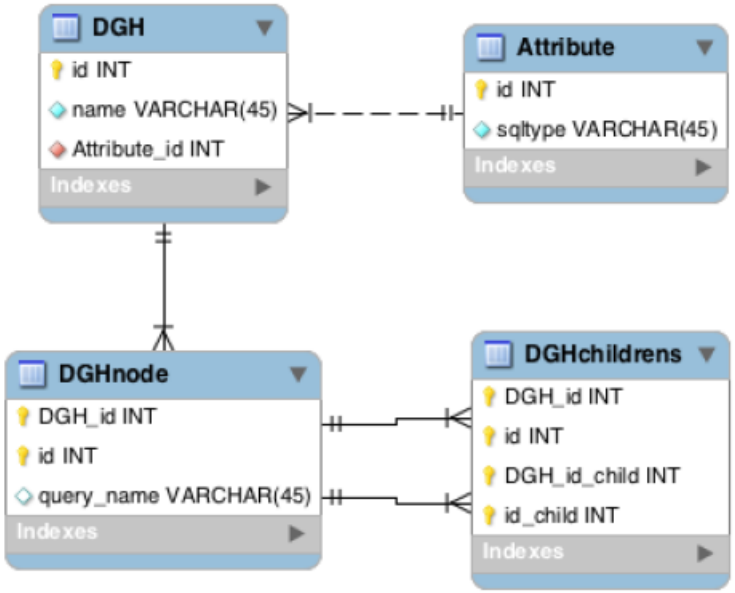
Anforderungen

- *automatische* Generierung… keine starren DGHs
- Speicherung generierter DGH in Datenbank
- Erweiterbarkeit für neue Clustering-Algorithmen / Ontologien

Konzept

Speicherung

- Datenbankschema von Martin Müller
- für MySQL-Server der Uni implementiert
- auch möglich auf anderen:
  - Postgres
  - MonetDB
  - DB2
  - …

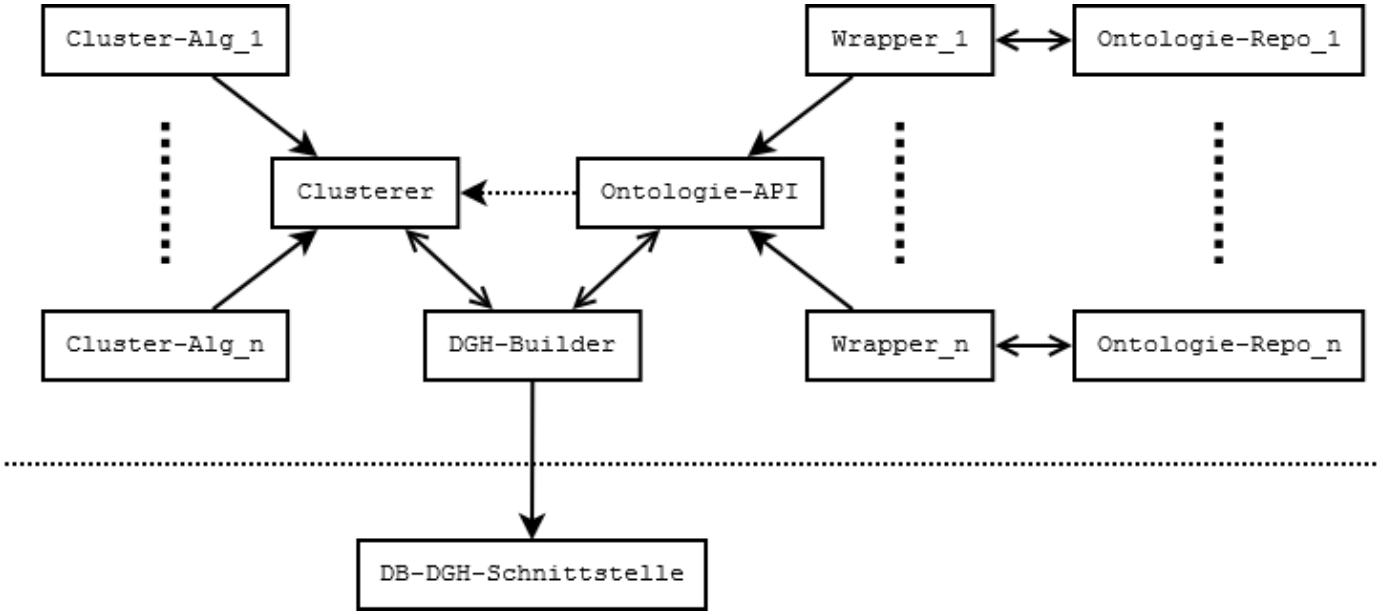


Quelle: Müller, Martin; *Technischer Bericht, CS-02-16*

Konzept

Grobentwurf

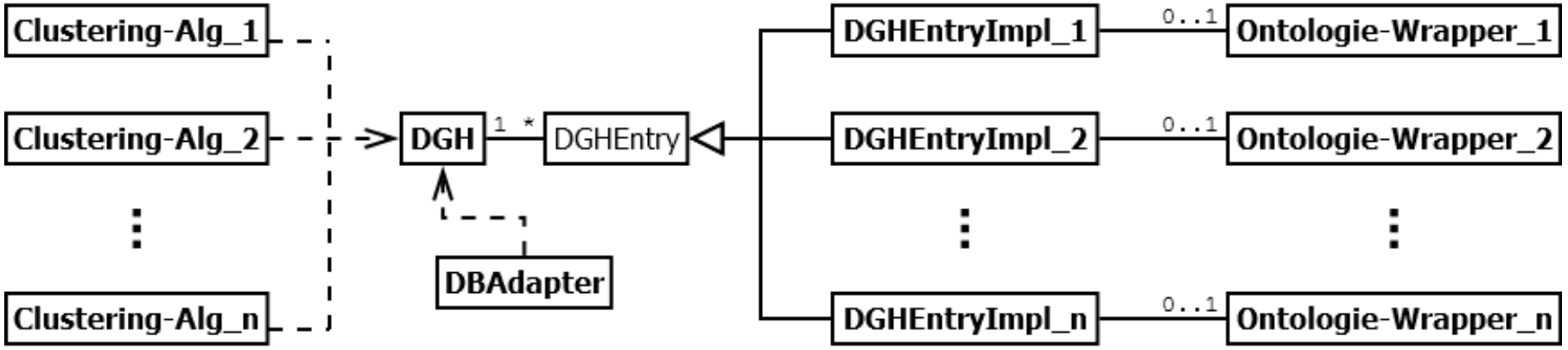
- Schnittstellen für Clustering-Algorithmen und Ontologien



Konzept

Neuer Grobentwurf

- Domänengeneralisierungshierarchie aus DGH-Objekt aufgebaut
- Clustering-Algorithmen arbeiten direkt auf DGH
- Ontologien in DGHEntry-Implementierung realisiert

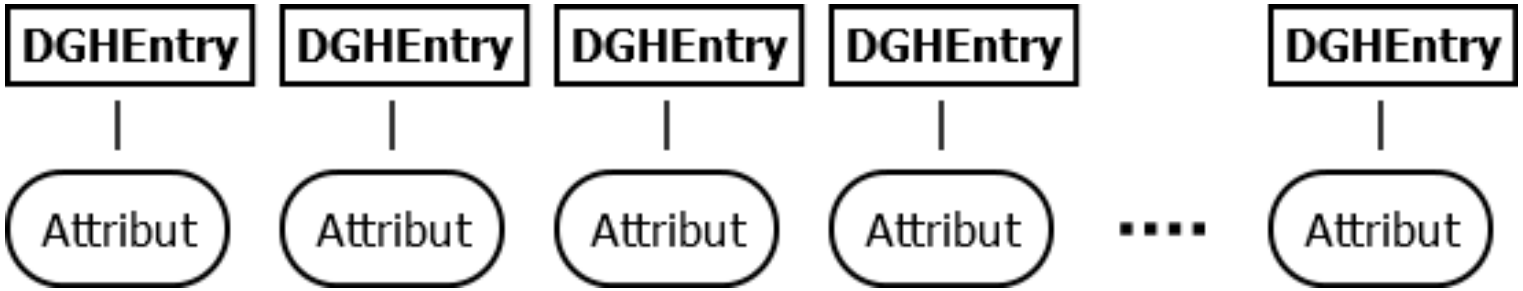


Konzept

DGH

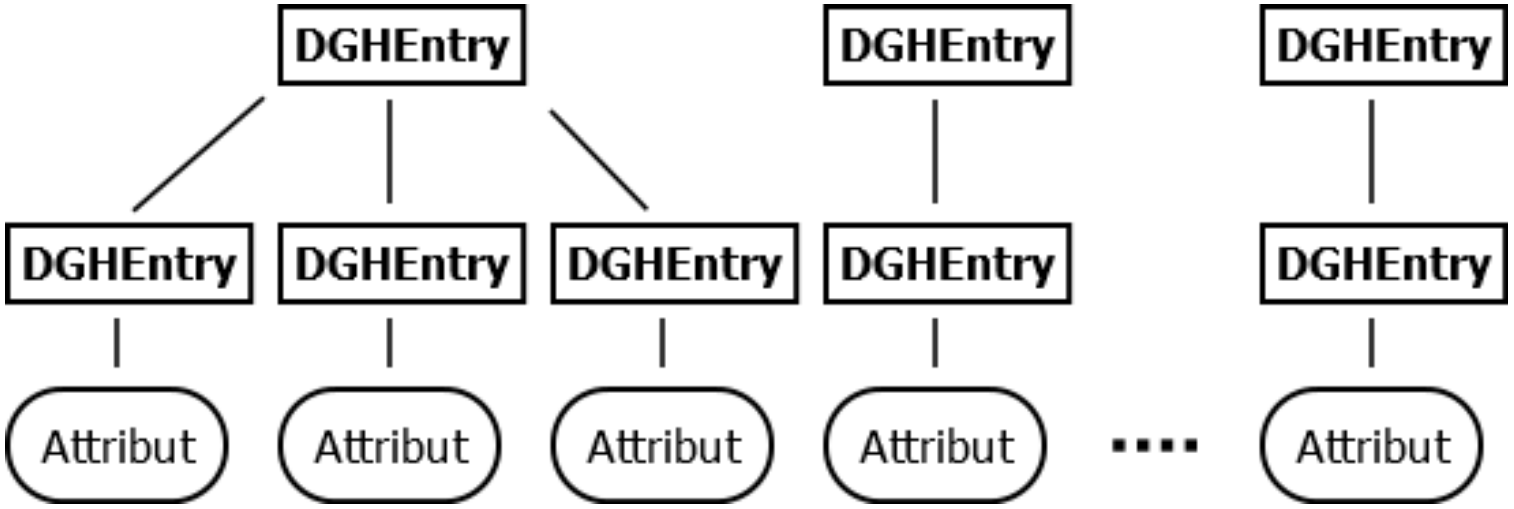
Konzept

DGH


DGHEntry
DGHEntry
DGHEntry
DGHEntry
DGHEntry
Attribut
Attribut
Attribut
Attribut
Attribut

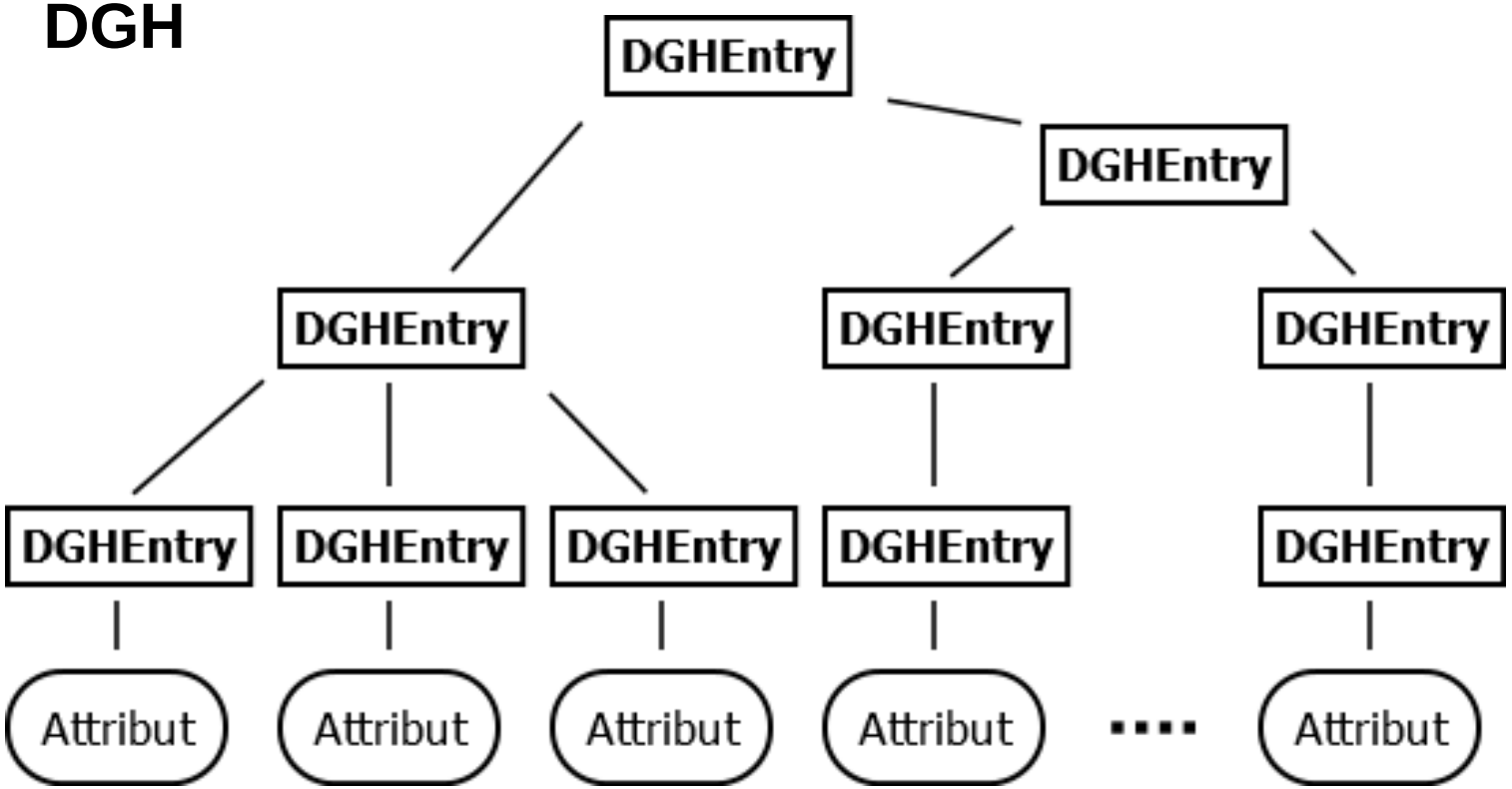
Konzept

DGH


DGHEntry
DGHEntry
DGHEntry
DGHEntry
DGHEntry
DGHEntry
DGHEntry
DGHEntry
Attribut
Attribut
Attribut
Attribut
Attribut

Konzept

DGH


DGHEntry
DGHEntry
DGHEntry
DGHEntry
DGHEntry
DGHEntry
DGHEntry
DGHEntry
DGHEntry
Attribut
Attribut
Attribut
Attribut
Attribut

Konzept

DGHEntry-Implementierungen

- **IntegerEntry** für Integer-Attribute
- **DateEntry** für Datums-Attribute
- **DoubleEntry** für Double-Attribute
- **ZIPEntry** für Postleitzahlen (numerische Implementierung)
- **LocationEntry** für Postleitzahlen (kategorische Implementierung)

Konzept

Clustering-Algorithmen

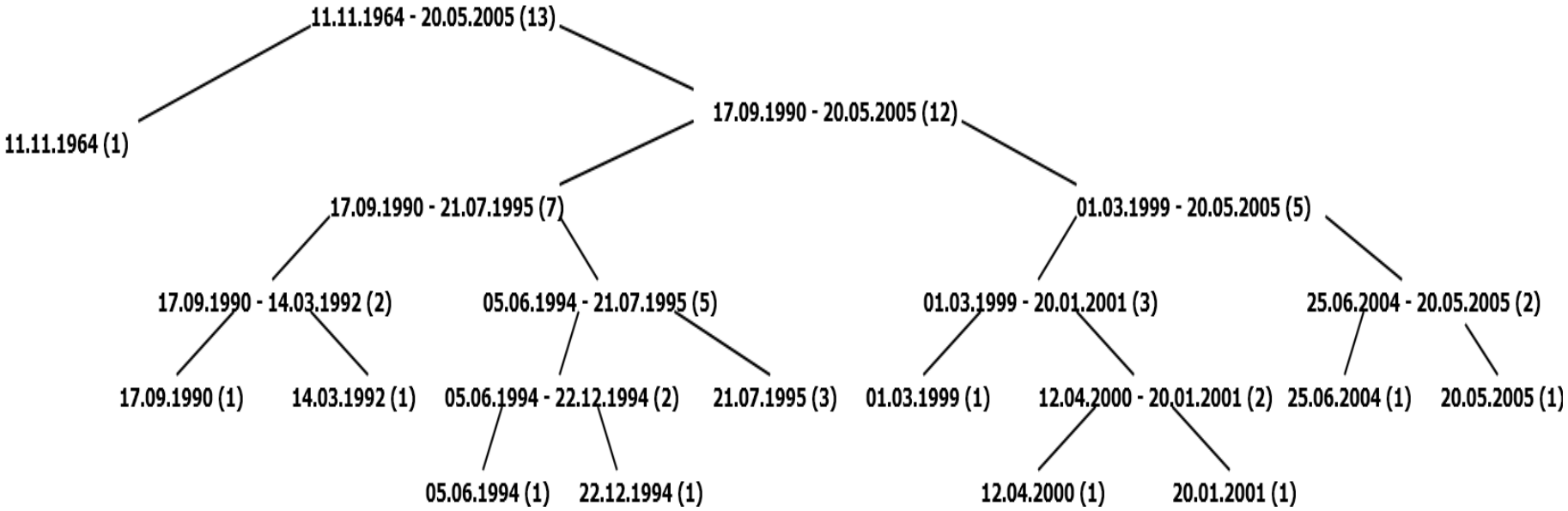
- hierarchisches Clustering
- gleichverteiltes Clustering

Ontologien

- OpenGeoDB

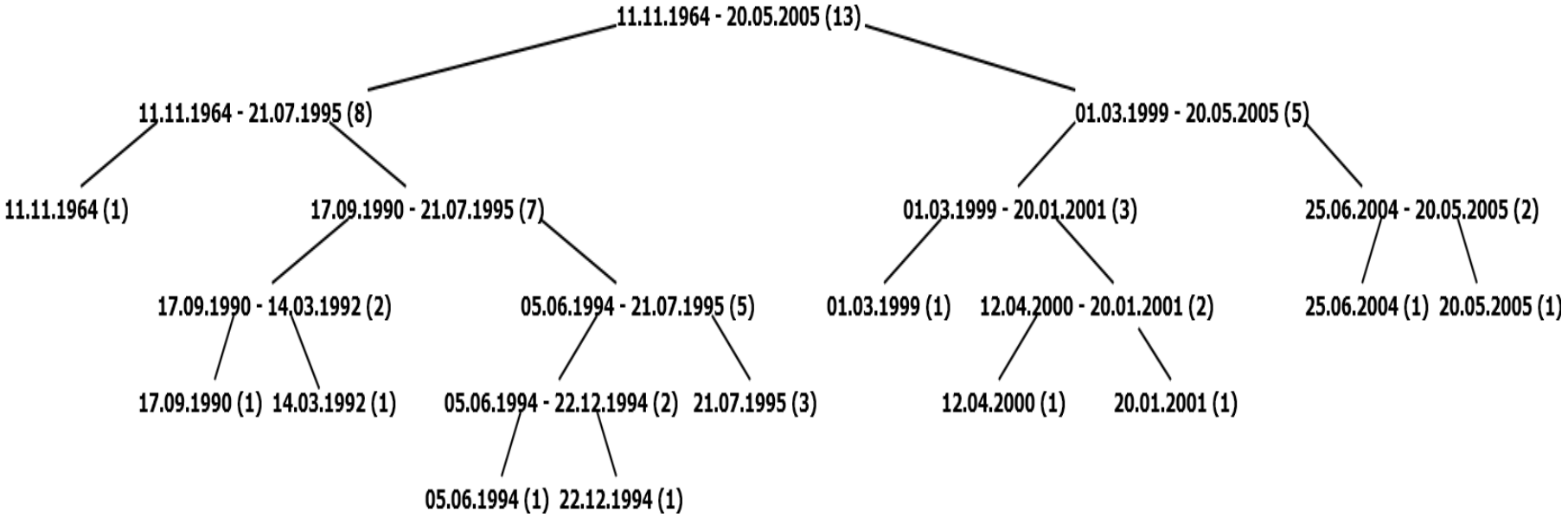
Ergebnisse

DateEntry & hierarchisches Clustering



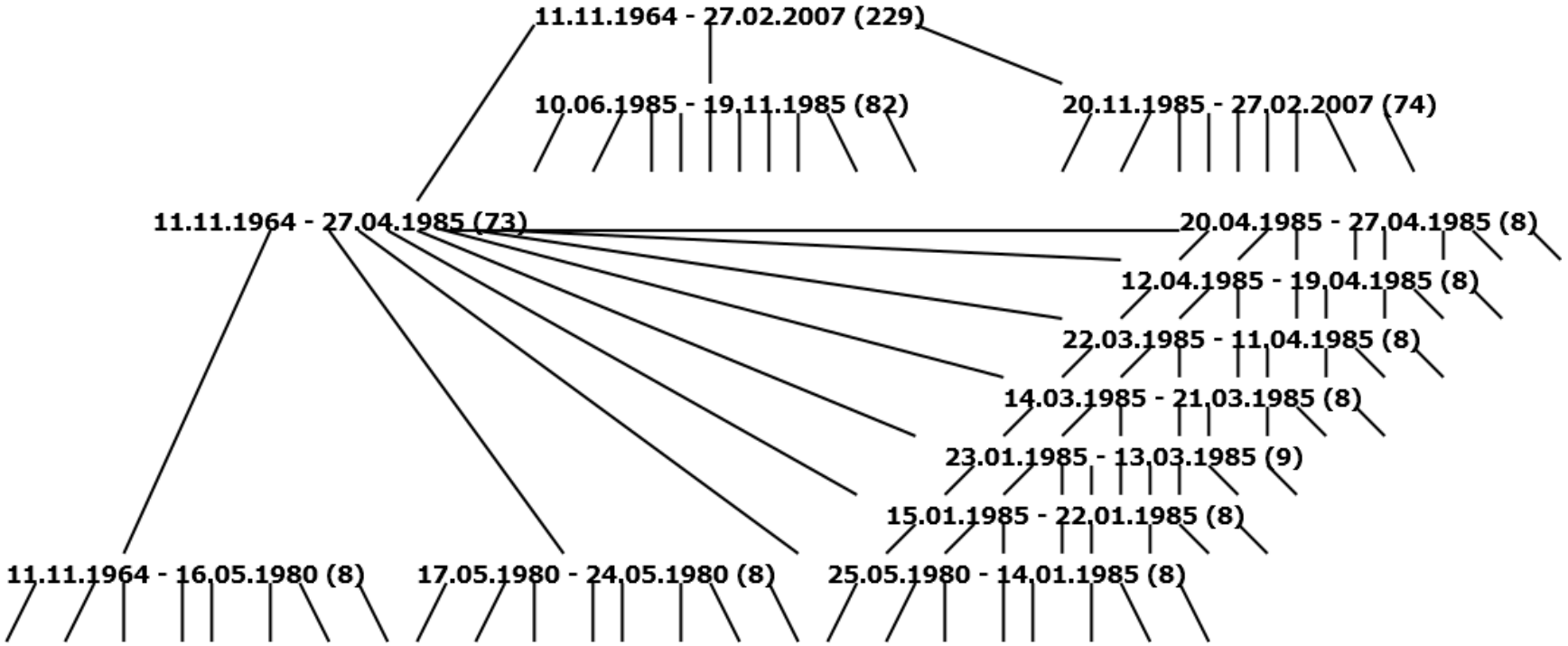
Ergebnisse

DateEntry & hierarchisches Clustering + Penalty



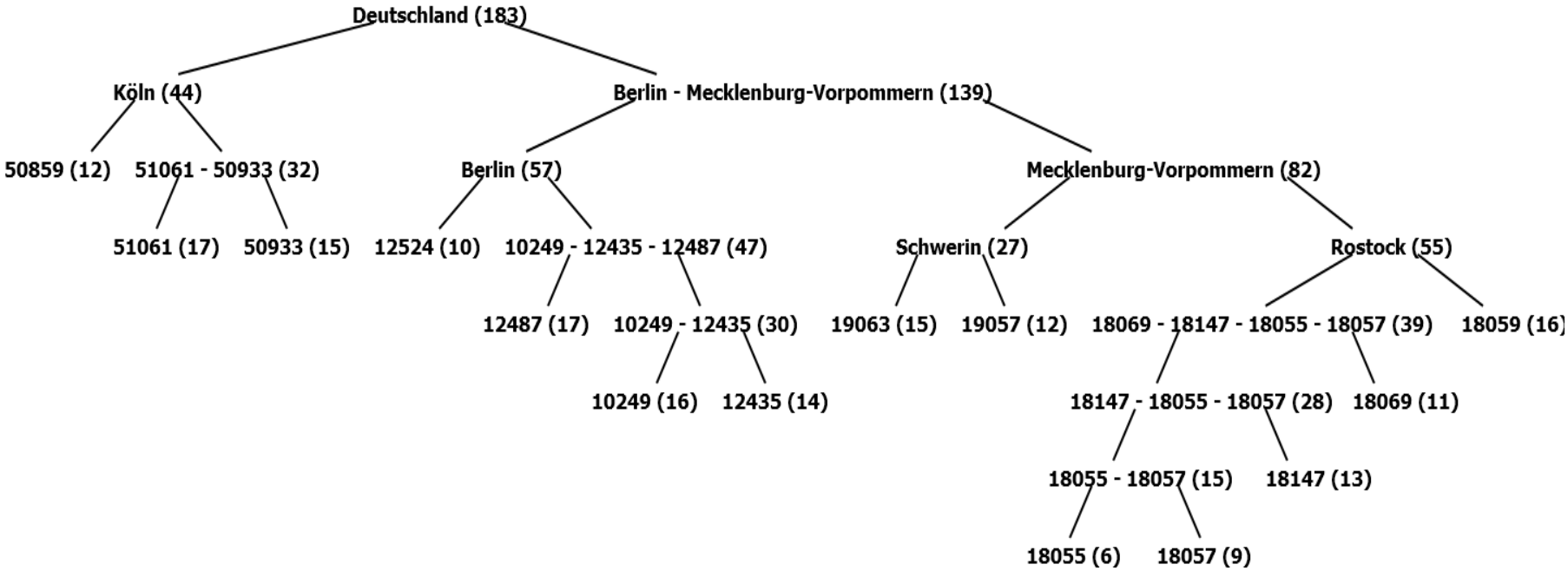
Ergebnisse

DateEntry & gleichverteiltes Clustering



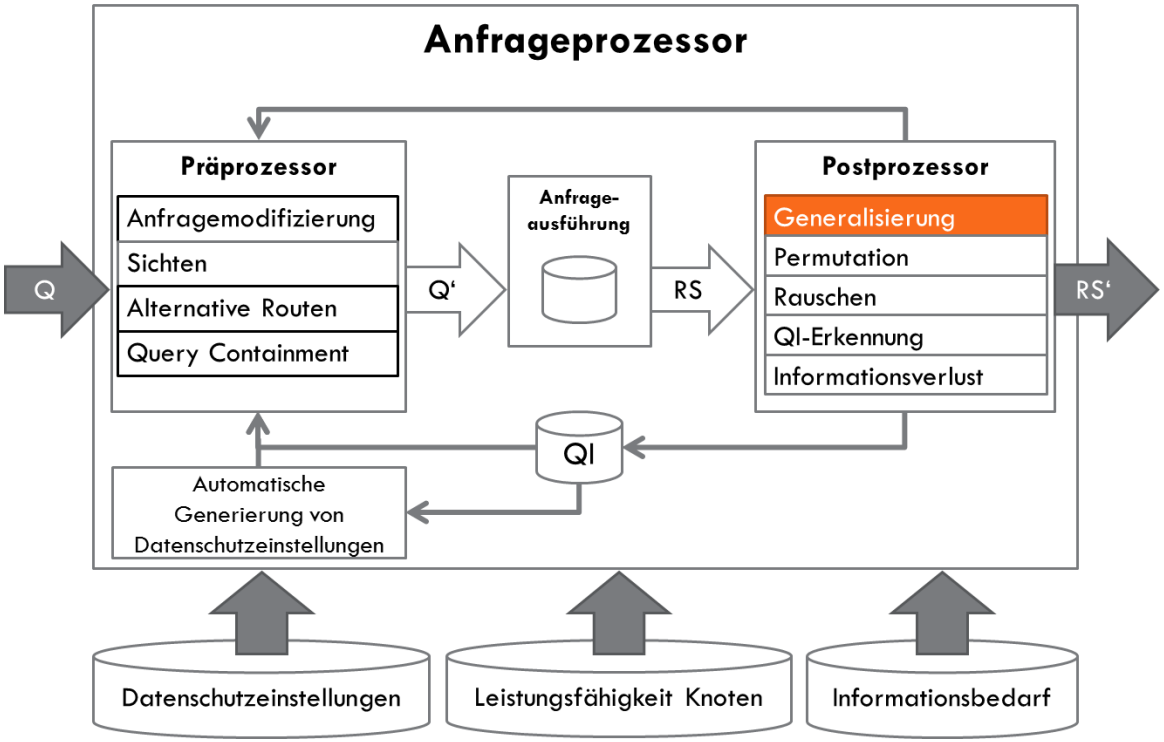
Ergebnisse

LocationEntry (OpenGeoDB) & hierarchisches Clustering



Einbindung

- in Postprozessor nach QI-Erkennung und vor Generalisierung



Zusammenfassung

- DGH-Generierung einfach für numerische Attribute
- mit Hilfe von Ontologien auch für kategorische Attribute
- Ontologien bieten Möglichkeit Werte zu substituieren

- gleichverteiltes Clustering bietet beste Ergebnisse
- hierarchisches Clustering in mehreren Fällen anwendbar

Vielen Dank für Ihre Aufmerksamkeit!