Automatische Generierung eines Mediators auf Basis einer deklarativen Spezifikation

Fachbereich Mathematik / Informatik
AG Künstliche Intelligenz

Kolloquium
Jörn Witte
09.02.2005

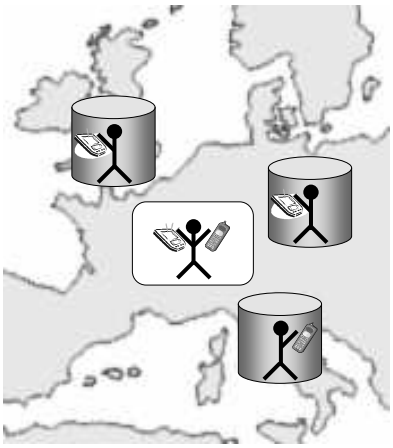
Überblick

- Einleitung
- Grundlagen
- Eigene Arbeit
  - Aufbau des Mediators
  - Generierung von Anfrageplänen
  - Festlegung einer Integrationsstrategie
- Zusammenfassung und Bewertung
- Ausblick

Überblick

- Einleitung
- Grundlagen
- Eigene Arbeit
  - Aufbau des Mediators
  - Generierung von Anfrageplänen
  - Festlegung einer Integrationsstrategie
- Zusammenfassung und Bewertung
- Ausblick

Motivation (I)



Ausgangslage:

- Gewachsene Systeme
- Zusammenhängende Daten über mehrere Datenquellen verteilt

Idee:

- Virtuelle Integration der heterogenen, verteilten und dynamischen Informationsquellen (Materialisierung in einer neuen Datenbank nicht notwendig)

Problematik:

- Strukturelle Heterogenitätskonflikte
- Semantische Heterogenitätskonflikte

Motivation (II)

Lösungsansätze zur virtuellen Integration von Daten:

- Schema Mediation in Peer Data Management Systems [Halevy03]
- MedMaker: A Mediation System Based on Declarative Specifications [Papakonstantinou96]
- Semantische Mediation [Wache03]

...da besondere Beachtung der semantischen Ebene

Motivation (III)

- Semantische Mediation nach [Wache03]
  - MeSA (**Me**CoTA **S**pecification **A**ssistent)
  - MeCoTA (**Me**diator with **Co**ntext **T**ransformation and **A**bstraction)

Auf kleinen Datenquellen getestet, aber:

- Theoretische Lösung - „proof-of-concept“
- Implementiert in Form eines Interpreters

...für große Datenquellen nicht unbedingt geeignet!

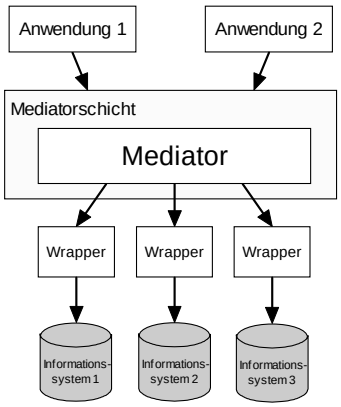
Ziele der Diplomarbeit

- Werkzeug zur Erzeugung eines spezifischen Mediators
- Optimierung des Anfrageprozesses
- Verbesserung der Interaktivität gegenüber MeCoTA

Überblick

- Einleitung
- Grundlagen
- Eigene Arbeit
  - Aufbau des Mediators
  - Generierung von Anfrageplänen
  - Festlegung einer Integrationsstrategie
- Zusammenfassung und Bewertung
- Ausblick

Was ist Mediation?



- ***Mediatoren*** "combine, integrate, and abstract the information" [Wiederhold92]
- Mediatoren realisieren eine ***Integrationsabbildung***
- ***Wrapper*** ermöglichen eine einheitliche Schnittstelle zu den heterogenen Systemen

Semantische Mediation nach [Wache03]

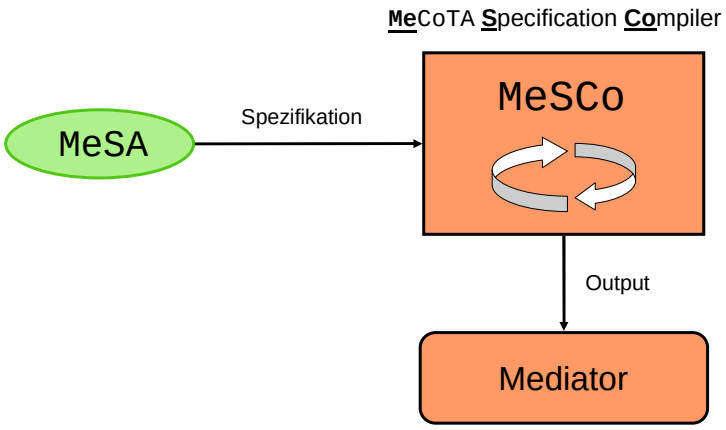
1. Akquisition der Vollständigen Beschreibung
   - syntaktische und semantische Beschreibung der Datenquellen
2. Identifikation semantisch äquivalenter Konzepte
3. Formulierung der Transformationsregeln
   - Definition von Anfragezerlegungsregeln
   - Definition von Kontexttransformationsregel (CT-Regel)

Assistenten unterstützen den Spezifikationsprozess

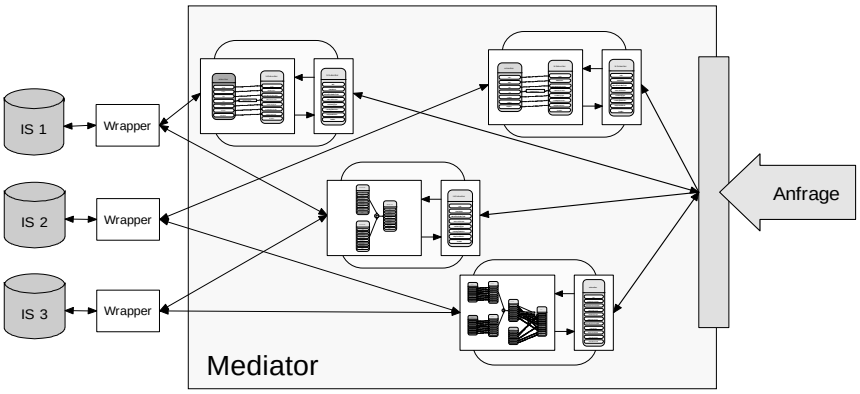
Überblick

- Einleitung
- Grundlagen
- Eigene Arbeit
  - Aufbau des Mediators
  - Generierung von Anfrageplänen
  - Festlegung einer Integrationsstrategie
- Zusammenfassung und Bewertung
- Ausblick

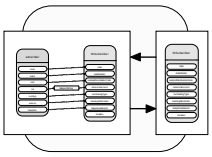
Einordnung dieser Arbeit



Aufbau des Mediators aus Integrationsstrukturen



Integrationsstruktur



Bestandteile der internen Repräsentation:

- Einheitliche Datenstruktur
  - Anfrageschnittstelle
  - Basierend auf dem syntaktischen Teil der Vollständigen Beschreibung [Wache03]
- Erweiterte Integrationsstruktur
  - Zerlegung einer Anfrage in Subanfragen an andere Informationssysteme
  - Beseitigung struktureller und semantischer Heterogenitätskonflikte
  - Grundlage ist ein bewiesener Anfrageplan

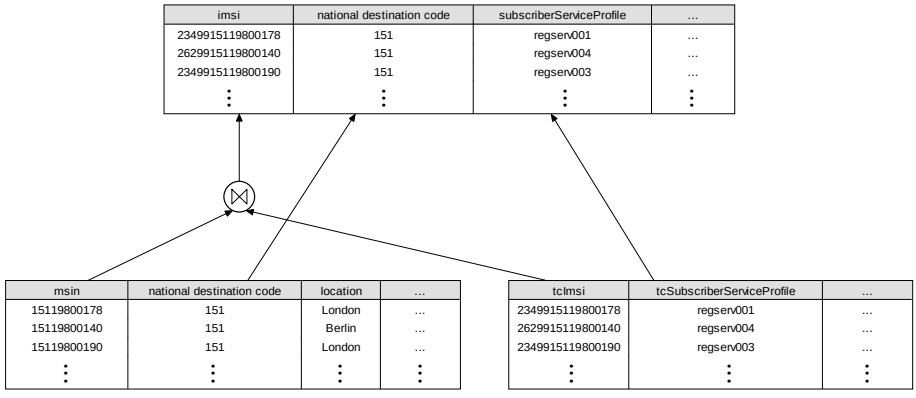
Umsetzung in einer Klassenstruktur

Anfrageplanung – Warum?

- *Der Ansatz von MeCoTA:*
  - MeCoTA wiederholt deduktives Beweisverfahren für jeden Datensatz
  - Blinde Suche über alle Kontexttransformationen für jeden Datensatz
- *Eigener Ansatz:*
  - Anfragepläne als Grundlage für den Aufbau der erweiterten Integrationsstruktur
- *Vorteile:*
  - Einmaliger Planungsprozess
  - Keine unnötige Suche zur Laufzeit

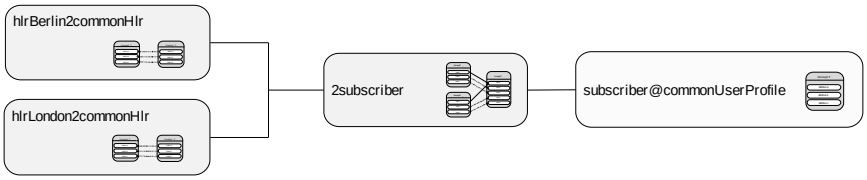
Erstellung eines
initialen Anfrageplans (I)
- Szenario -

- *Grundlage des initialen Anfrageplans:*
  - Anfragezerlegungsregel



Erstellung eines initialen Anfrageplans (II)

- Zerlegung einer Anfrage in Subanfragen [Xiaolei96]
- Auflösung struktureller Integrationskonflikte
- Grundgerüst für den bewiesenen Anfrageplan



Semantische Auswertung einer Anfragezerlegungsregel (I)

Ziel:

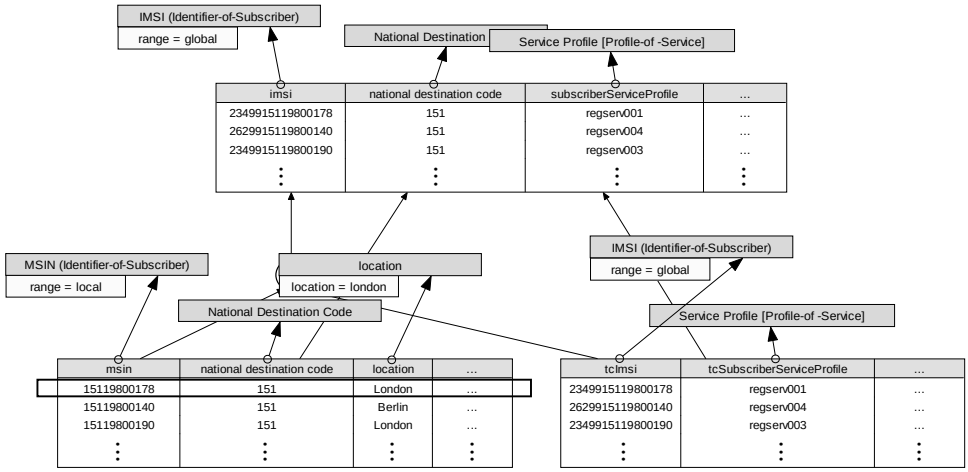
- Semantische Heterogenitätskonflikte aufdecken
- Sequenzen von Kontexttransformationen zur Konfliktbeseitigung für einzelne Konzeptkorrespondenzen ermitteln

Problematik:

- Verschiedene Lösungswege in Abhängigkeit von den Daten

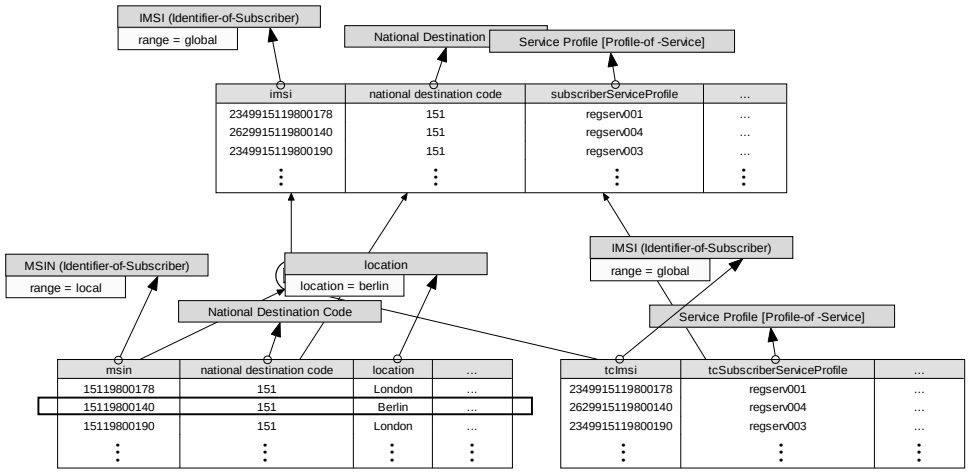
Semantische Auswertung einer Anfragezerlegungsregel (I)

- Szenario -



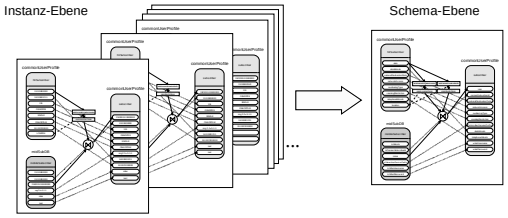
Semantische Auswertung einer Anfragezerlegungsregel (I)

- Szenario -



Semantische Auswertung einer Anfragezerlegungsregel (II)

- Formalisierung in Hornlogik
- Logische Beweisverfahren auf Schema-Ebene
  - Verwendung der SLD-Resolution in Verbindung mit klassischer Unifikation [Robinson79]
- Im Gegensatz dazu führt MeCoTA Beweisverfahren auf Instanz-Ebene durch
  - Verwendung der CT-Resolution und des CT-Kalküls [Wache03]



Semantische Auswertung einer Anfragezerlegungsregel (III)

Idee:

- Zielkonzepte aus den korrespondierenden Quellkonzepten einer Anfragezerlegungsregel herleiten
- Iteratives Beweisverfahren auf Basis der SLD-Resolution
  - dabei *spezielle* CT-Regeln sukzessive aus der Wissensbasis entfernen
  - erfolgreiche Ableitungsketten als Auswertungsgrundlage verwenden
- Abhängigkeiten von Daten sowie die Operationen der CT-Regeln werden erst zur Laufzeit berücksichtigt

Anfragezerlegungsregel gilt als bewiesen, wenn alle Konzeptkorrespondenzen erfolgreich abgeleitet werden konnten

Erstellung eines bewiesenen Anfrageplans

- *Grundlage:*
  - Initialer Anfrageplan
  - Ergebnisse der semantischen Auswertungen der Anfragezerlegungsregeln
- Beinhaltet das Wissen zur Auflösung struktureller und semantischer Heterogenitätskonflikte
- Dient als Basis für die erweiterte Integrationsstruktur

Integrationsstrategie

- *Pipelining*
  - Parallelisierung der einzelnen Teilschritte
  - Daten werden von einem Prozess zum nächsten weitergereicht
  - Keine Vorteile bei blockierenden Operationen

  Vorteile:
  - Hohe Interaktivität
  - Keine Beschränkung der zu verarbeitenden Datenmenge

Überblick

- Einleitung
- Grundlagen
- Eigene Arbeit
  - Aufbau des Mediators
  - Generierung von Anfrageplänen
  - Festlegung einer Integrationsstrategie
- Zusammenfassung und Bewertung
- Ausblick

Zusammenfassung und Bewertung (I)

- Generierter Mediator besteht aus einer Menge von Integrationsstrukturen in Form von Java-Klassen
- Anfragepläne dienen dabei als Modellierungsgrundlage für den Mediator
- Anfrageplangenerierung in drei Schritten:
  1. Initialer Anfrageplan
  2. Semantische Auswertung der Anfragezerlegungsregeln
  3. Bewiesener Anfrageplan
- Pipelining als Integrationsstrategie

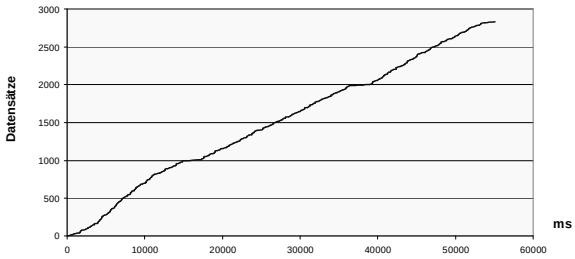
Zusammenfassung und Bewertung (II)

- Domänenspezifischer, autarker Mediator
- Integrationsstruktur
  - Uniforme Anfrageschnittstelle
  - Abbildung des integrativen Wissens
- Unterstützung einer Konkretisierung der semantischen Beschreibung in Abhängigkeit von den Daten
- Interaktivität gegenüber MeCoTA wurde verbessert

Zusammenfassung und Bewertung (III)

Ergebnisse:

- geringe Startzeit, d.h. frühe Bereitstellung integrierter Ergebnisse im Vergleich zu MeCoTA
- Ausgeglichene Pipeline durch konstante Zufuhr von Informationen
- Ausgeglichene Pipeline ist vorteilhaft für die Gesamtzeit




```
...
MeSCoCriteria crit = new MeSCoCriteria();
crit.add(SubscriberPeer.NATIONALDESTINATIONCODE,"151");

SourceThread cup = new CommonuserprofileSubscriber
              CommonuserprofileSubscriberSourceThread();

OutputTable t = new OutputTable(cup);
t.startQuery(crit);

...
```

Überblick

- Einleitung
- Grundlagen
- Eigene Arbeit
  - Aufbau des Mediators
  - Generierung von Anfrageplänen
  - Festlegung einer Integrationsstrategie
- Zusammenfassung und Bewertung
- Ausblick

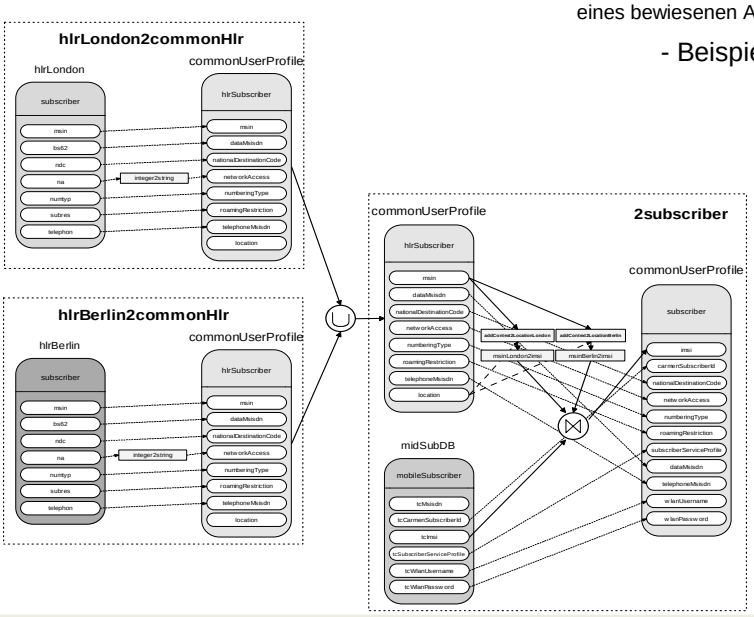
Ausblick

- Umkehrung der Kontexttransformation
  - Ermöglicht die Propagierung von Bedingungen an die Subanfragen
- Optimierung der Anfragepläne
  - z.B. Veränderung der Join-Reihenfolge, interessante Sortierungen
- Verfahren zur Steigerung der Informationsqualität
  - z.B. Duplikaterkennung

Anhang

Grafische Darstellung
eines bewiesenen Anfrageplans

- Beispiel -



Überführung einer Anfragezerlegungsregel

- Beispiel -

```
'2subscriber'&&
template('subscriber',VAR_LAB_SUB,complex,VAR_VAL_SUB)::[
            'imsi' -->> template('imsi',
                                 VAR_LAB_IMSI,
                                 string,
                                 VAR_VAL_IMSI)::[]@'commonUserProfile',
...
]@'commonUserProfile'
<<--
[template('hlrSubscriber',VAR_LAB_SUB,complex,VAR_VAL_SUB)::[
            'msin' --*> template('msin',
                                 VAR_LAB_MSIN,
                                 string,
                                 VAR_VAL_MSIN)::[]@'commonUserProfile',
...
]@'hlrBerlin'].
```

Anzahl der einzelnen Beweise bzw. unterschiedlicher Wissensbasen ist abhängig

- von der Anzahl der Attributkonzepte
- sowie der Kardinalität der Anfragezerlegungsregel

```
concept(VAR_VAL_MSIN,msin,string,commonUserProfile)
term(VAR_VAL_MSIN,MSIN [Identifier of Subscriber])
context(VAR_VAL_MSIN,range,local)

query(imsi) ← concept(VAR_VAL_IMSI,imsi,string,commonUserProfile)
            ∧ term(VAR_VAL_IMSI,IMSI [Identifier of Subscriber])
            ∧ context(VAR_VAL_IMSI,range,global)
```


Überführung einer Kontexttransformationsregel

- Beispiel -

```
'addContext2LocationLondon'&&
template(VAR_NAME,
         'Location Identifier'::['location' **>> 'london'],
         string,
         'London')::[]@'commonUserProfile'
<<**
[??template(VAR_NAME,
            'Location Identifier'::[],
            string,
            'London')::[]@'commonUserProfile'].
```

Abhängigkeiten von Daten sowie die Operationen der CT-Regeln werden erst zur Laufzeit berücksichtigt

```
context(VAR_NAME@addContext2LocationLondon, location, london)
← concept(VAR_NAME@addContext2LocationLondon, VAR_NAME, string, commonUserProfile)
  ∧ term(VAR_NAME@addContext2LocationLondon, Location Identifier)
```