Semantische Datenintegration

Methoden und Architekturen

Jakob Voß

Hochschule Hannover

2017-06-10

Grundlagen

Semantische Datenintegration

Semantische Datenintegration ist die sinnvolle Zusammenführung von uneinheitlichen Daten verschiedener Quellen.

Semantische Datenintegration: Arten

Semantische Datenintegration ist die sinnvolle Zusammenführung von uneinheitlichen Daten verschiedener Quellen.

1. Data Wrangling
2. Integriertes Informationssystem
3. Föderiertes Informationssystem

Arten der (semantischen) Datenintegration

- Einfache Datenintegration (Data Wrangling)
  - Hand- und Kopfarbeit
  - Einzelfall
  - $\Rightarrow$ *Datenbereinigung & Archivierung...*
- Integriertes Informationssystem
  - Fasst mehrere Datenquellen zusammen
  - Definiert ein integriertes Modell
  - Ermöglicht Anfragen oder erzeugt regelmäßig Ausgaben
  - $\Rightarrow$ *Mashup, ETL, Metasuche...*
- Föderiertes Informationssystem
  - Fasst mehrere autonome Datenquellen zusammen
  - Basiert auf Abmachungen zwischen Datenquellen
  - $\Rightarrow$ *WWW, Semantic Web...*

Herausforderungen

- Verteilte Datenquellen
- Autonome Datenquellen
- Uneinheitliche Daten
  - Unterschiedliche Formate
  - Unterschiedliche Modelle
  - Unterschiedliche Semantik

Architekturen

- Virtuelle Integration:
  Integration zur Anfrage, einheitliche Sicht
  - Mashup, Metasuche...
- Materiele Integration:
  Integration vor Anfrage, einheitliche Daten
  - ETL, Data Warehousing...

Beispiele und Übungen

Beispiel: Discovery-Systeme

- Metasuche
- Föderierte Suche
- Integrierte Suche

Übung: Föderierte Informationssintegration

SPARQL Federated Queries

- RDF als gemeinsame Datenstrukturierungssprache
- SPARQL als gemeinsame Abfragesprache
- hoffentlich gemeinsame Datenschemata
  - gemeinsame Ontologien
  - gleiche Daten und -Modelle
  - gleiche Identifier (idealerweise URIs)

SPARQL Federated Query

- Wikidata-SPARQL Endpoint:
  `http://query.wikidata.org/`
- Nomisma-SPARQL Endpoint:
  `http://nomisma.org/query`

SPARQL Federated Query

```
SELECT * WHERE {

  # ...Wikidata...

  SERVICE <http://nomisma.org/query> {

    # ...Nomisma...

  }
}
```

Sinnvolle Abfragen?

SPARQL Federated Query

```
SELECT * WHERE {
  ?item wdt:P31 wd:Q41207 ;
        wdt:P2950 ?nid
  BIND( uri(concat('http://nomisma.org/id/', ?nid))
        as ?nomismaID )
  SERVICE <http://nomisma.org/query> {
    ?nomismaID skos:definition ?def .
  }
} LIMIT 10
```

Beispiel von `https://www.wikidata.org/wiki/User:Smalyshev_(WMF)/Federation`

Herausforderungen und Methoden

Herausforderungen

- Fehlende oder Fehlerhafte Daten
- Unterschiedliche Identifier
- Unterschiedliche Datenmodelle
- Technik (Antwortzeiten etc.)

Einige Schwierigkeiten

- Fehlende oder Fehlerhafte Daten
  - Informationsqualität
- Unterschiedliche Identifier
  - Objektidentifikation und Duplikaterkennung
- Unterschiedliche Datenmodelle
  - ($\rightarrow$) Mapping und Matching
- Technik (Antwortzeiten etc.)
  - Hier nicht Thema

Methoden

- Datenbereinigung
- Duplikaterkennung
- Datenfusion
- Informationsqualität

Datenbereinigung

- Datenfehler erkennen
- Datenfehler lösen

Datenfehler erkennen

Umfrage

Datenfehler erkennen: Arten von Fehlern

- Schemafehler
  - Unzulässige Werte
  - Ad-Hoc Schema
  - Inkonsistente Angaben
- Datenfehler
  - Fehlende Werte
  - Tippfehler
  - Dummy-Eingaben (-, `XXXX`, `0`...)
  - Falsche Werte (veraltete, unbekannt...)
  - Zeile/Spalte/Feld/.. verwechselt
  - Duplikate

Datenfehler lösen

- Messen, Testen, Validieren
- Data Lineage
- Zusätzliche Constraints
- Statistische Auswertungen
- Normalisieren, Formatieren, Referenzieren

kein Feedback, keine Qualität

Duplikaterkennung

- (Semi)automatische Erkennung verschiedener Datensätze die sich auf das gleiche Objekt beziehen
- Ziel: Identifier und/oder normalisierte Daten
- Aufwändige Vergleiche ($n \times n/2 - n$)
- Precision & Recall

Verfahren zur Duplikaterkennung

- Edit-Distance
  - Ähnlichkeit von Zeichenketten
- Tupel-Ähnlichkeit
  - $\Rightarrow$ Information Retrieval (z.B. *tf-idf*)
- Gruppierung/Partionierung
- Hashwert
  - Abbildung auf kleinere Datenmenge

Zusammenfassung: Datenintegration

- Einmalig / Integriert / Föderiert
- Virtuell / Materiell
- Datenfehler erkennen und damit Leben
- Duplikaterkennung an verschiedenen Stellen

Literatur und Quellen

Quellen dieser Folien: https://github.com/hshdb/MWM-317-02/