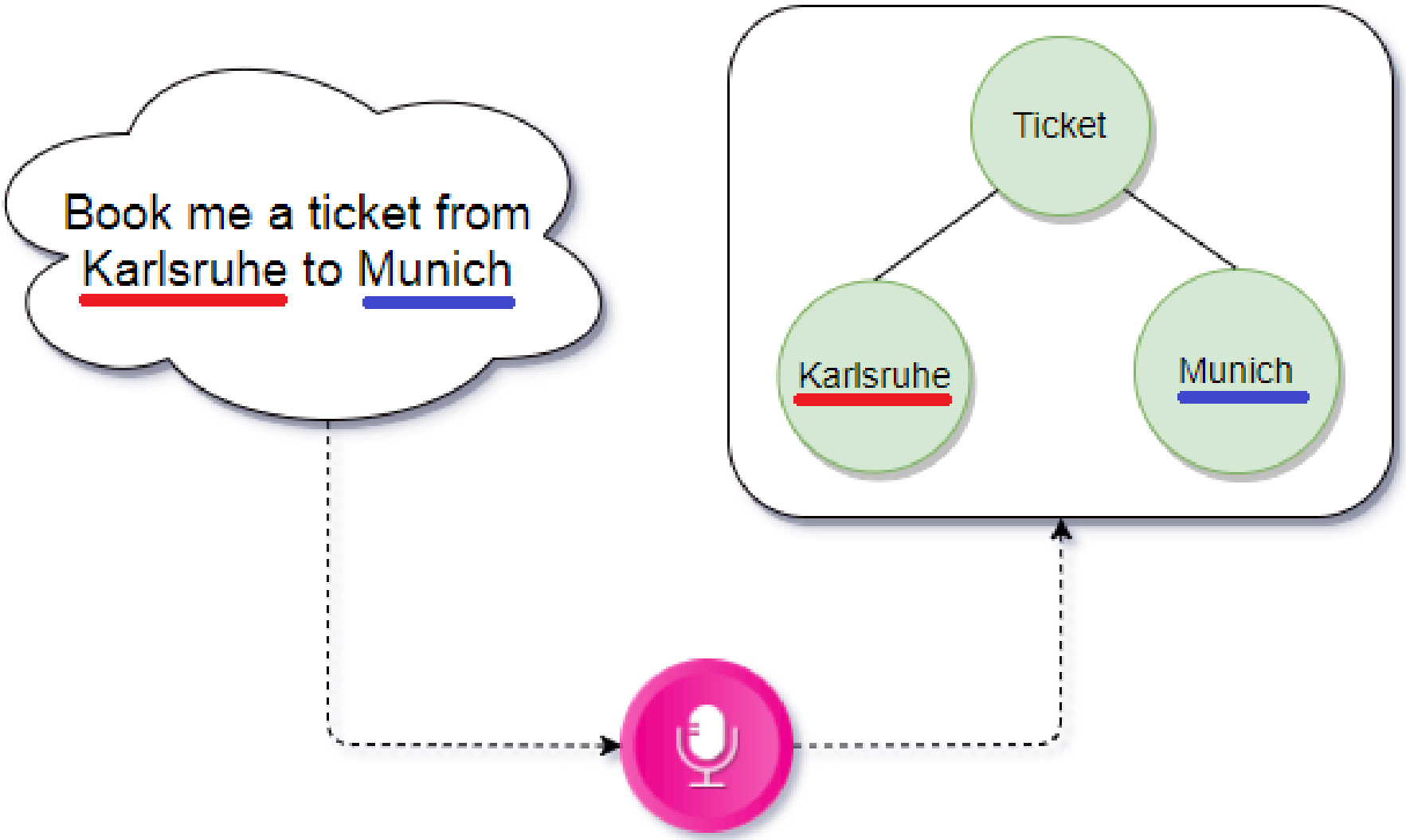
Karlsruher Institut für Technologie

Bachelorarbeit: Wissensbasierte Identifikation von Wertebereichen einer aktiven Ontologie

Yauhen Makhotsin, betreut von Martin Blersch

IPD Tichy, Fakultät für Informatik

Motivation

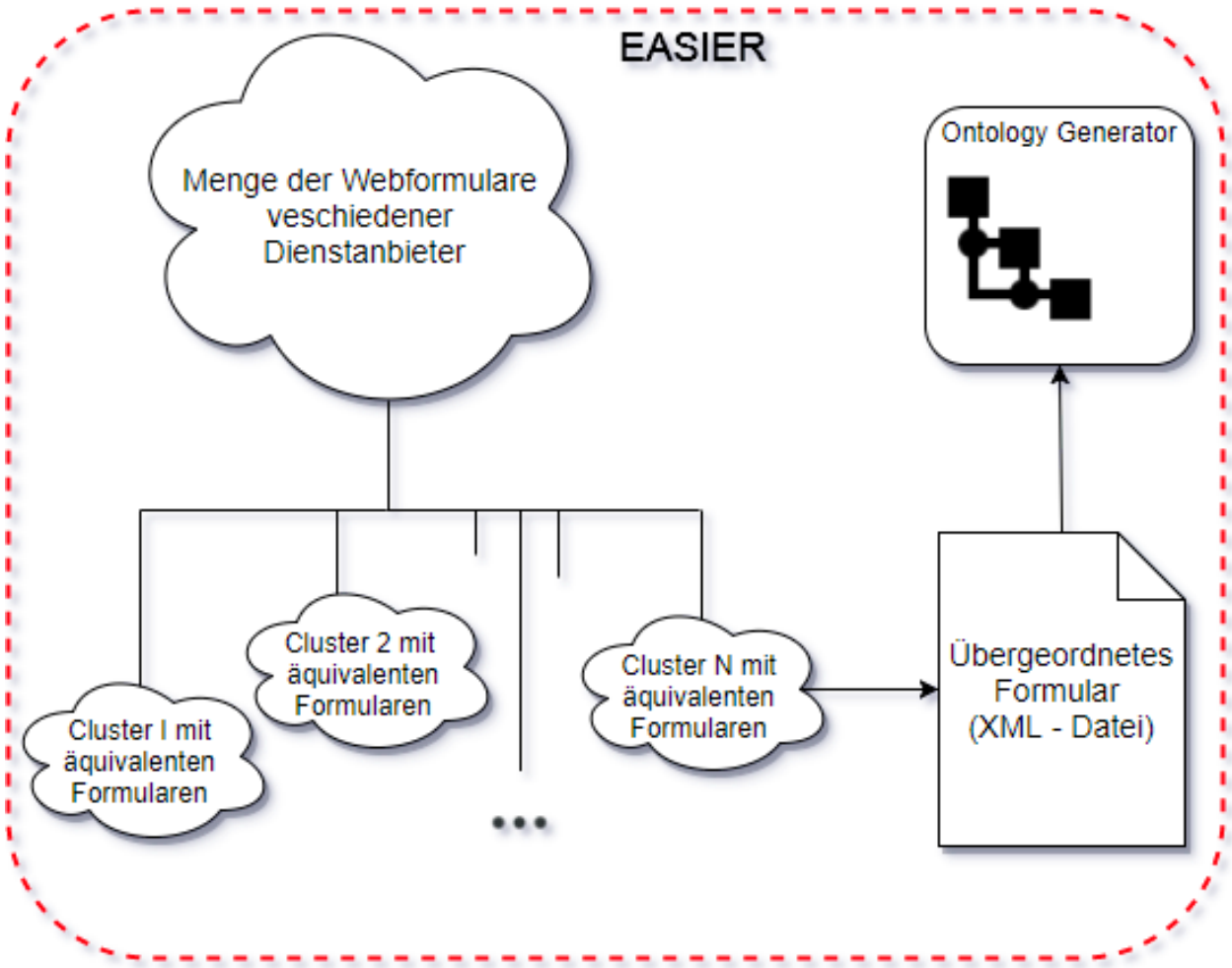

Book me a ticket from Karlsruhe to Munich
Ticket
Karlsruhe
Munich

Motivation

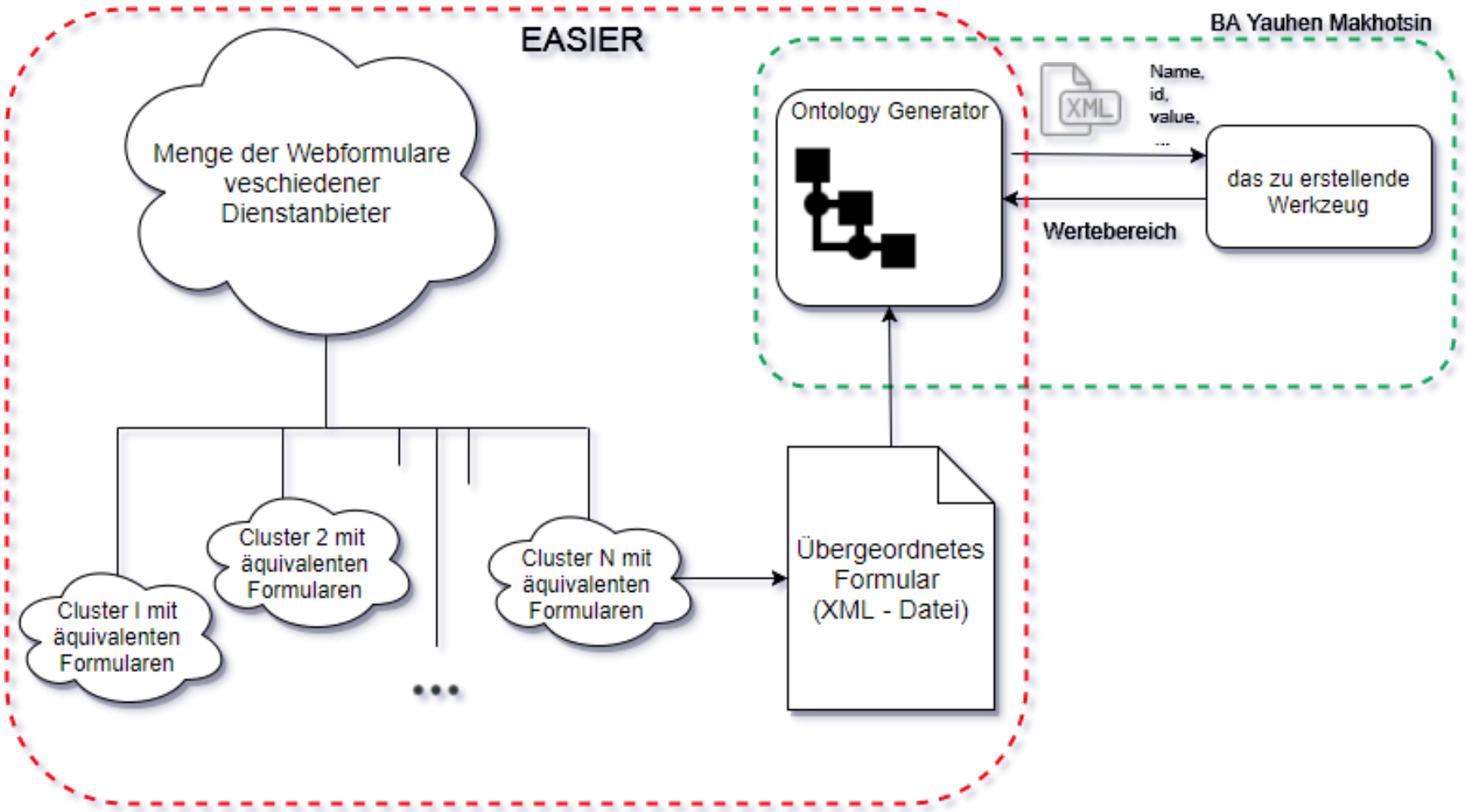
Motivation

Book me a ticket to New York

Motivation


EASIER
Menge der Webformulare veschiedener Dienstanbieter
Ontology Generator
Cluster I mit äquivalenten Formularen
Cluster 2 mit äquivalenten Formularen
Cluster N mit äquivalenten Formularen
Übergeordnetes Formular (XML - Datei)

Ziel der Arbeit



Verwandte Arbeiten

Unsupervised Domain Ontology Learning from Text [Gee17]

- Suche nach Textkorpora
- Extraktion von Termen
- Extraktion von taxonomischen Relationen
- Extraktion von nicht taxonomischen Relationen
- Ontologieerstellung

Verwandte Arbeiten

Erkennung von Hyperonymie:

- Bedingungen:
  - **t0** Suffix von **t1**
  - beide Domänenterme
- Folgerung:
  - **t0** Oberbegriff von **t1**
- Beispiel:
  - “polysaccharide” Oberbegriff von “homopolysaccharide”

Verwandte Arbeiten

Erkennung von Hyperonymie:

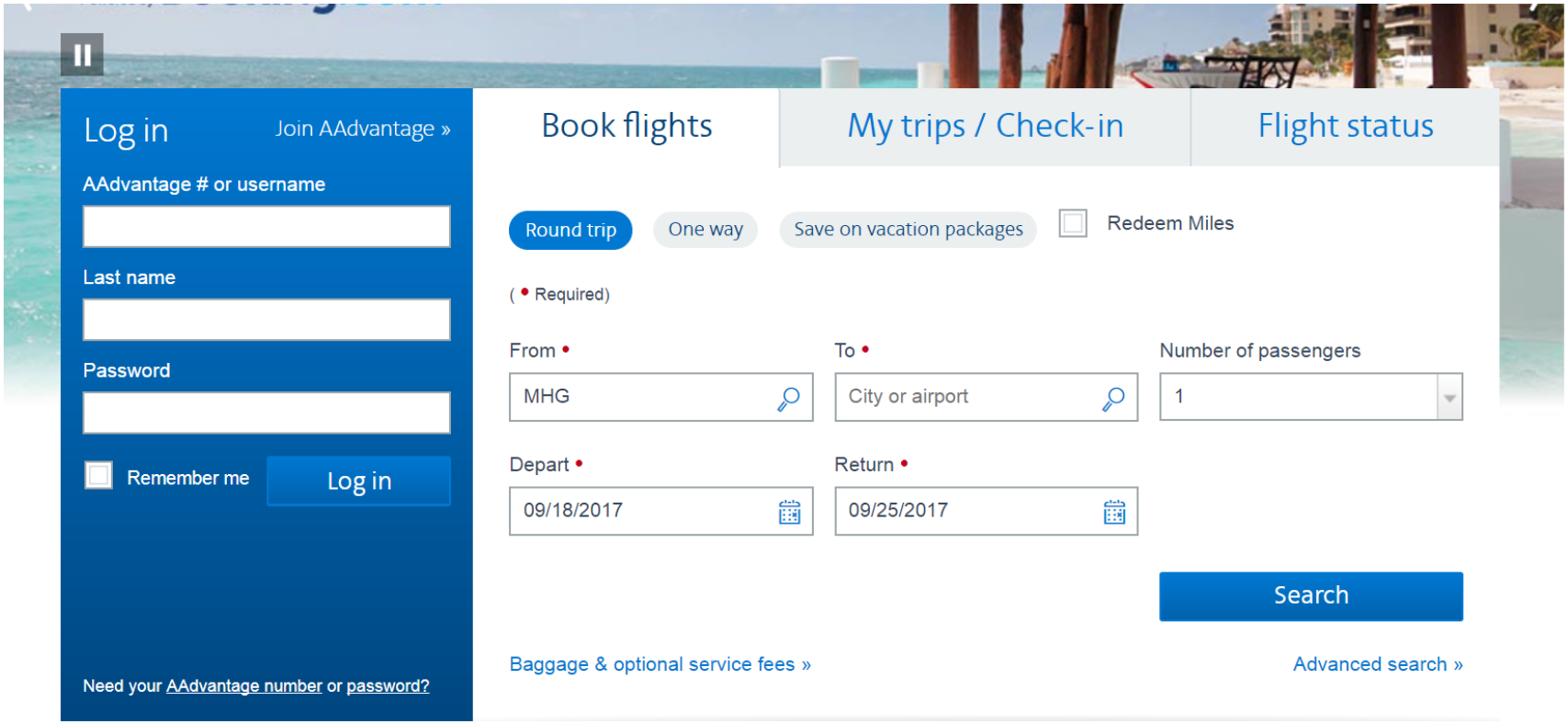
- Bedingungen:
  - **t0** ist Kopfterm von **t1**
  - einer ist Domänenterm
- Folgerung:
  - **t0** Oberbegriff von **t1**
- Beispiel:
  - “corn” Oberbegriff von “sweet corn”

Verwandte Arbeiten

Building a Domain Knowledge Base from Wikipedia: a Semi-supervised Approach [**Chen2016BuildingAD**]

- Entdeckung der Domänenkonzepte
  - Auswahl der domänenrelevanten Hashtags von Stackoverflow
  - Vorbereitung der Hashtags
  - Abbildung auf Artikel in Wikipedia
  - Entdeckung weitere Konzepte mithilfe der Kategorien in Wikipedia
- Entdeckung der Relationen
  - Durch semantische Verarbeitung
  - Falls ein Artikel einer Kategorie gehört und beide als Konzepte erkannt, wird eine “subclass of” - Relation erstellt.

Semantik in HTML



Semantik in HTML

```
<input type="text" name="originAirport" value="MHG" id="
    reservationFlightSearchForm.originAirport" class="
    aaAutoComplete" placeholder="City␣or␣airport" >
```

Semantik in HTML

```
<input type="text" name="originAirport" value="MHG" id="
    reservationFlightSearchForm.originAirport" class="
    aaAutoComplete" placeholder="City or airport" >
```

Semantik in HTML

```
<input type="text" name="originAirport" value="MHG" id="
    reservationFlightSearchForm.originAirport" class="
    aaAutoComplete" placeholder="City or airport" >
```

Semantik in HTML

```
<input type="text" name="originAirport" value="MHG" id="
    reservationFlightSearchForm.originAirport" class="
    aaAutoComplete" placeholder="City␣or␣airport" >
```

Semantik in HTML

```
<input type="text" name="originAirport" value="MHG" id="
    reservationFlightSearchForm.originAirport" class="
    aaAutoComplete" placeholder="City␣or␣airport" >
```

Semantik in HTML

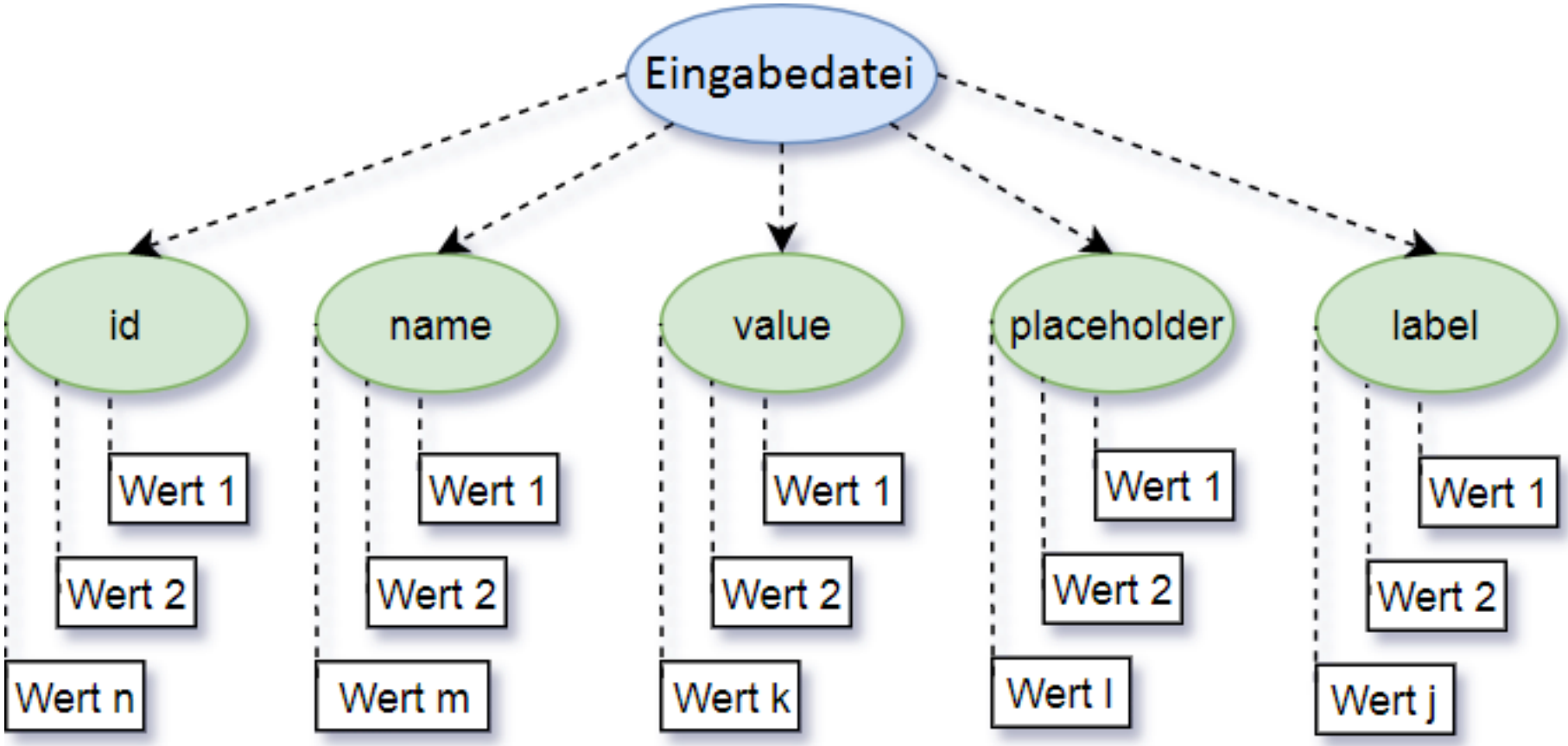
- Im Formular enthalten:
  1. Oberbegriffe: **City**, **Airport**
  2. Domänenwert: **MHG**
- Zu untersuchen:
  1. Welche Unterbegriffe von **City** und **Airport** gibt es?
  2. Was ist **MHG**?
  3. Welche ähnliche Begriffe zu **MHG** gibt es?

Semantik in HTML

Identifikation der Domäne:

1. Klassifikation der übermittelten Attributwerte in mögliche Oberbegriffe und Domänenwerte
2. Filterung der resultierenden Listen
3. Befragung der externen Datenquellen
4. Konsolidierung der Ergebnisse

Eingabe


Eingabedatei
id
name
value
placeholder
label
Wert 1
Wert 2
Wert n
Wert 1
Wert 2
Wert m
Wert 1
Wert 2
Wert k
Wert 1
Wert 2
Wert l
Wert 1
Wert 2
Wert j

Verarbeitung der Eingabe

Klassifizierung der Attributwerte:

- Klasse von der Attributart abhängig
- Zuweisung einer Bewertung
- Beispiel
  - Der Wert des Labels ist ein **Oberbegriff**
  - Der Wert des Elements **value** ist entweder ein **Oberbegriff** oder ein **Domänenwert**

Verarbeitung der Eingabe

Tabelle: Wahrscheinlichkeiten des Auftrenens der Oberbegriffe und Domänenwerte

Parameter	P(Oberbegriff)	P(Domänenwert)	P(irrelevant)
label	0.4	0	0,6
value	0.25	0.75	0
placeholder	0.46	0.08	0,46
name	0.35	0	0,65
id	0.43	0	0,57

Verarbeitung der Eingabe

Präfixe:

- Suche nach Präfixen
- Das Präfix wird gelöscht
- Der Wert wird in die richtige Liste verschoben
- Bewertung erhöhen
- Beispiel:
  - “Enter the city” ——> “city” ist ein Oberbegriff
  - “Example: Mannheim” ——> “Mannheim” ist ein Domänenwert

Verarbeitung der Eingabe

Aufteilung in Teilbegriffe:

- Nach Großschreibung. Z.B:
  - originCity ——> {origin, city, origin city}
  - airportName ——> {airport, name, airport name}
- Nach Trennzeichen. Z.B:
  - desired location ——> {desired, location, desired location}
  - arrival_time ——> {arrival, time, arrival time}
- Abschließende Entfernung der Stoppwörter wie:
  - Attributnamen (z.B. Label, Placeholder, ID usw)
  - Artikel

Verarbeitung der Eingabe

Lexikalische Analyse:

- Eigennamen ermitteln
- Substantive ermitteln
- Zusammensetzungen mit Substantiven ermitteln
- Rest löschen

Verarbeitung der Eingabe

Ergebnis:

- Liste der möglichen Oberbegriffe
- Liste der möglichen Domänenwerte

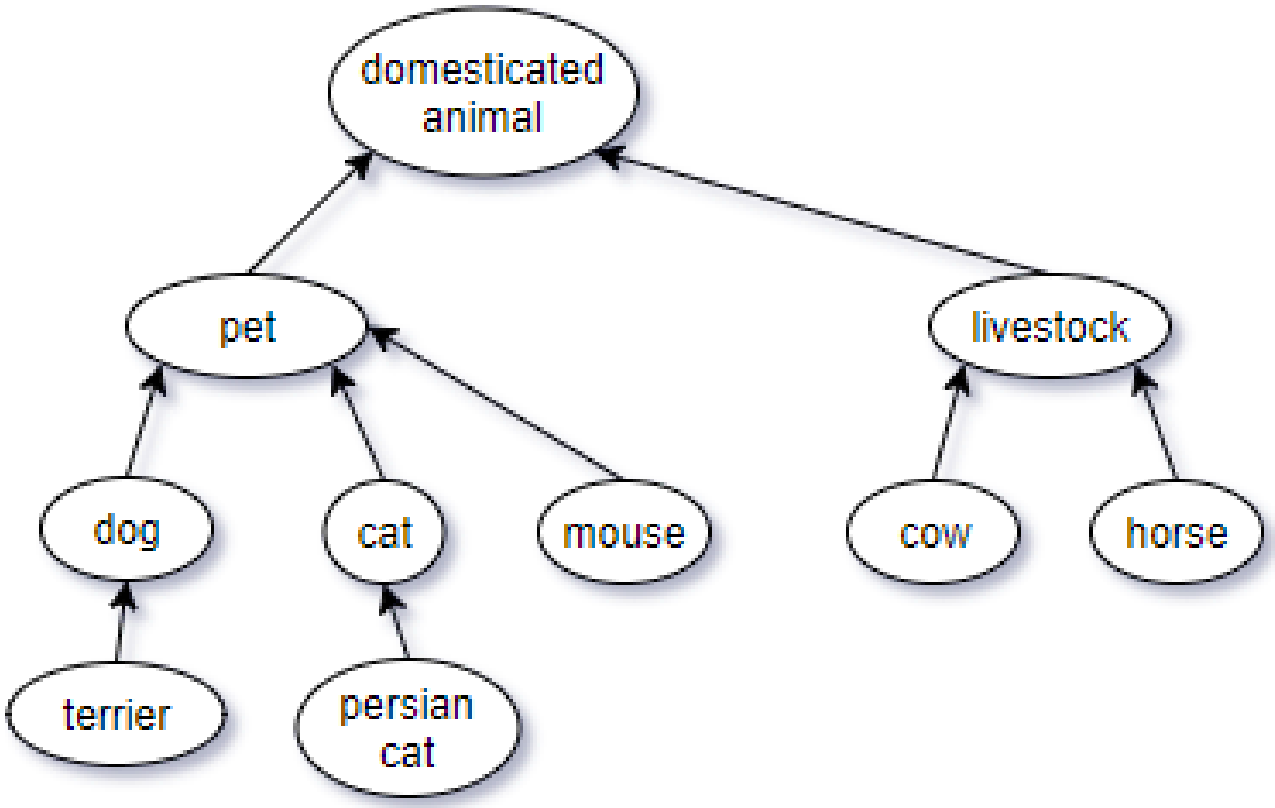
Befragung der externen Datenquellen

Befragte Datenquellen:

- WordNet
- Wikipedia
- ResearchCyc

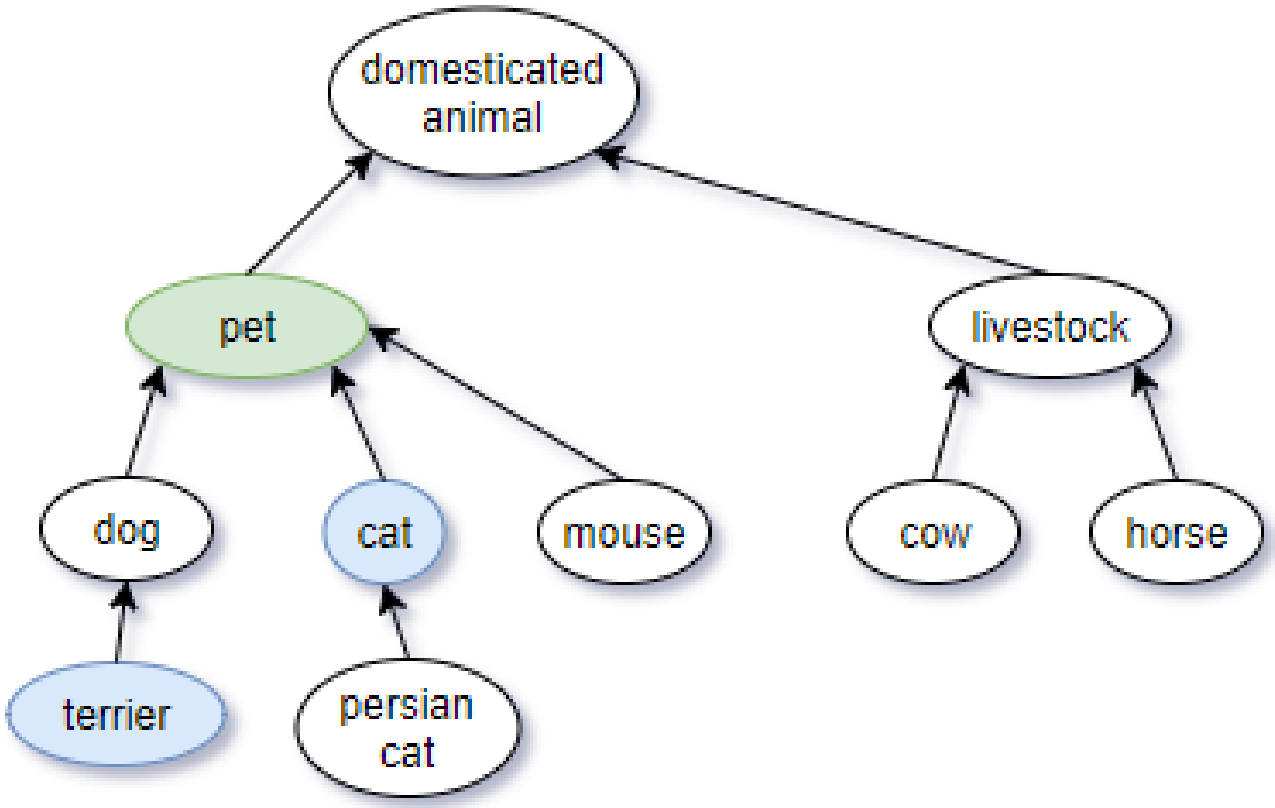
Befragung der externen Datenquellen

WordNet:



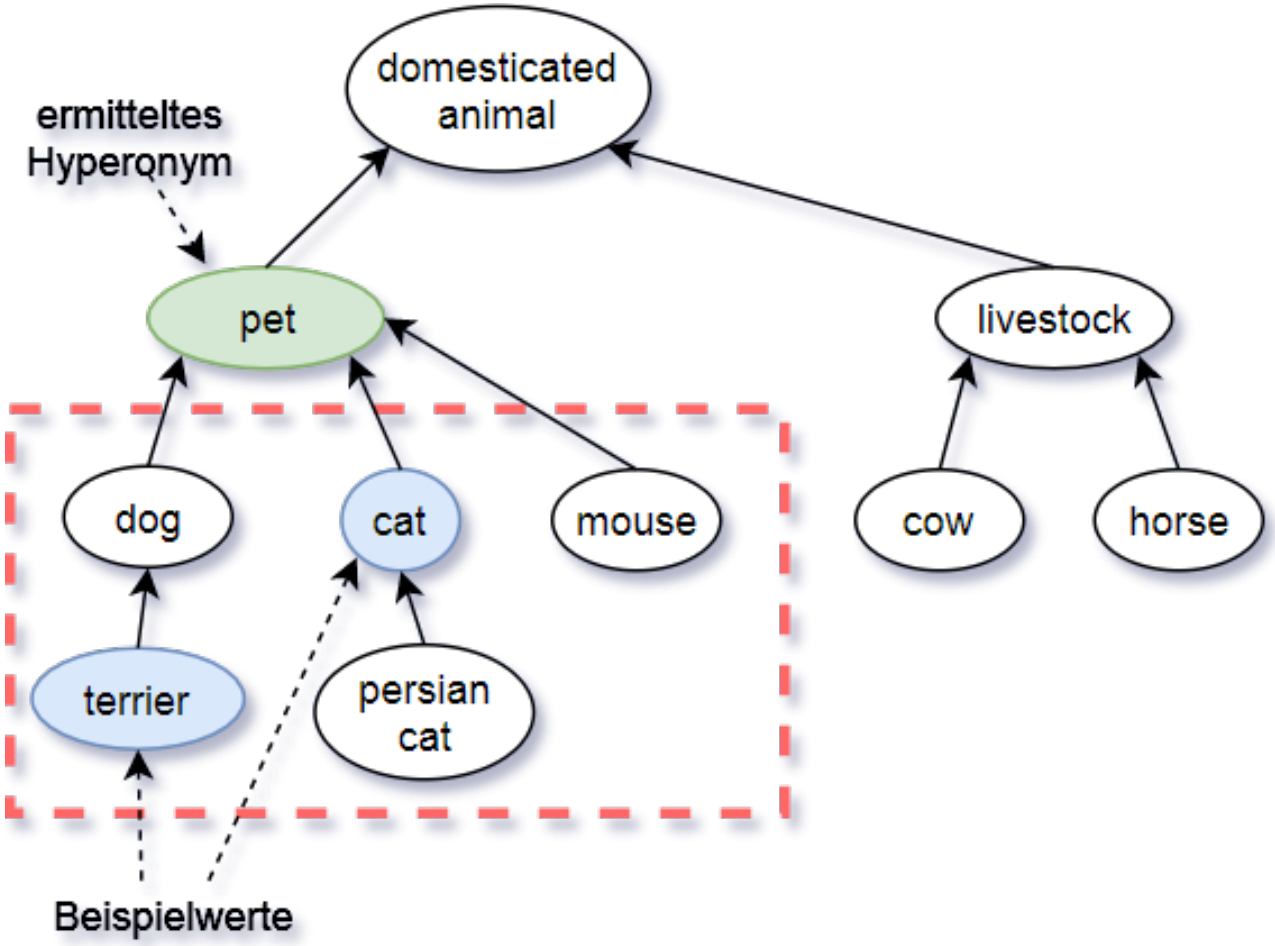
Befragung der externen Datenquellen

WordNet:



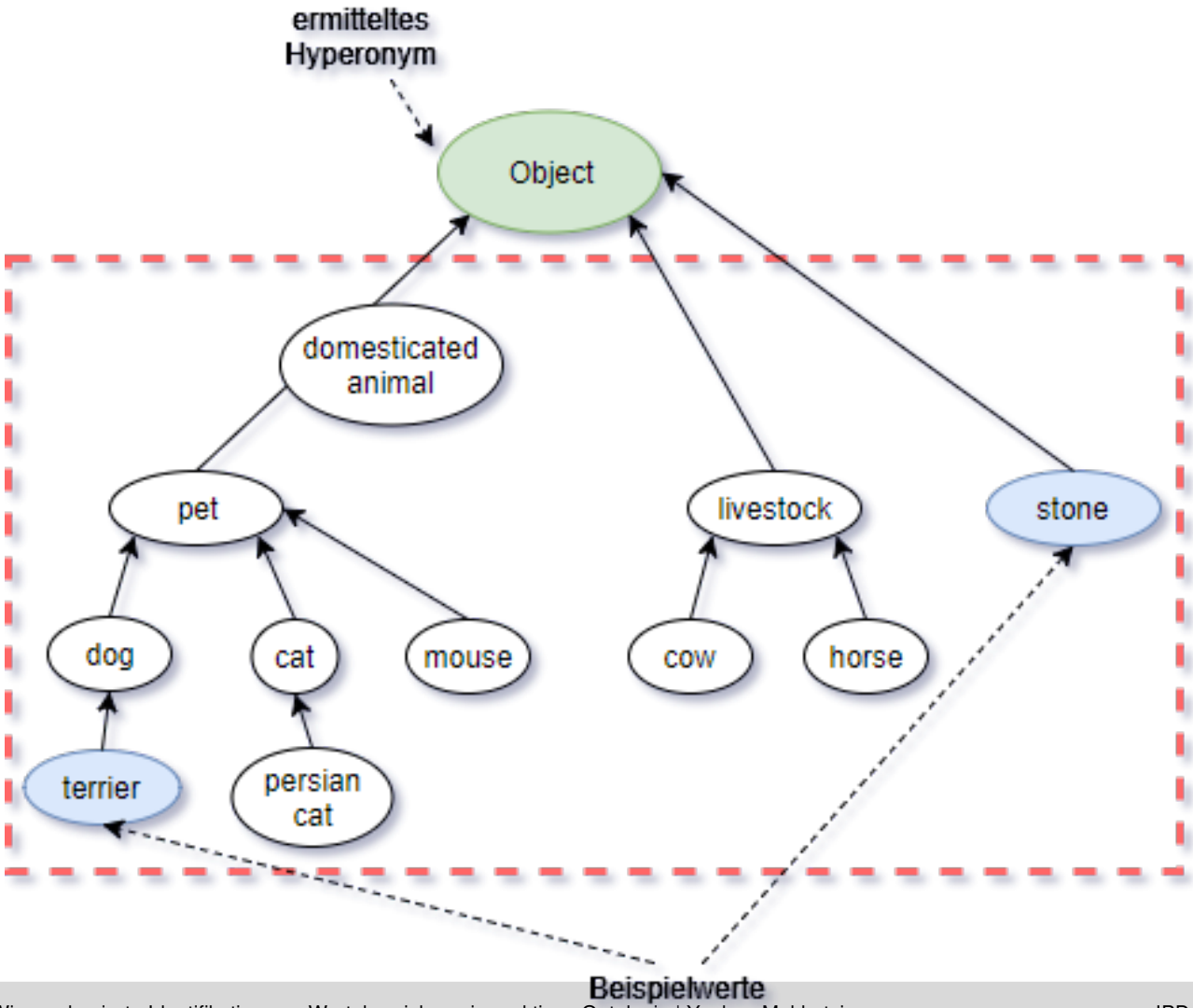
Befragung der externen Datenquellen

WordNet:



Befragung der externen Datenquellen

WordNet:



Befragung der externen Datenquellen

Semantische Ähnichkeit (Similarity):

- Nur semantisch ähnliche Werte verwenden
- Similarity(Auto, Fahrrad) > Similarity(Auto, Baum)
- Semantische Ähnlichkeit nach Wu-Palmer:

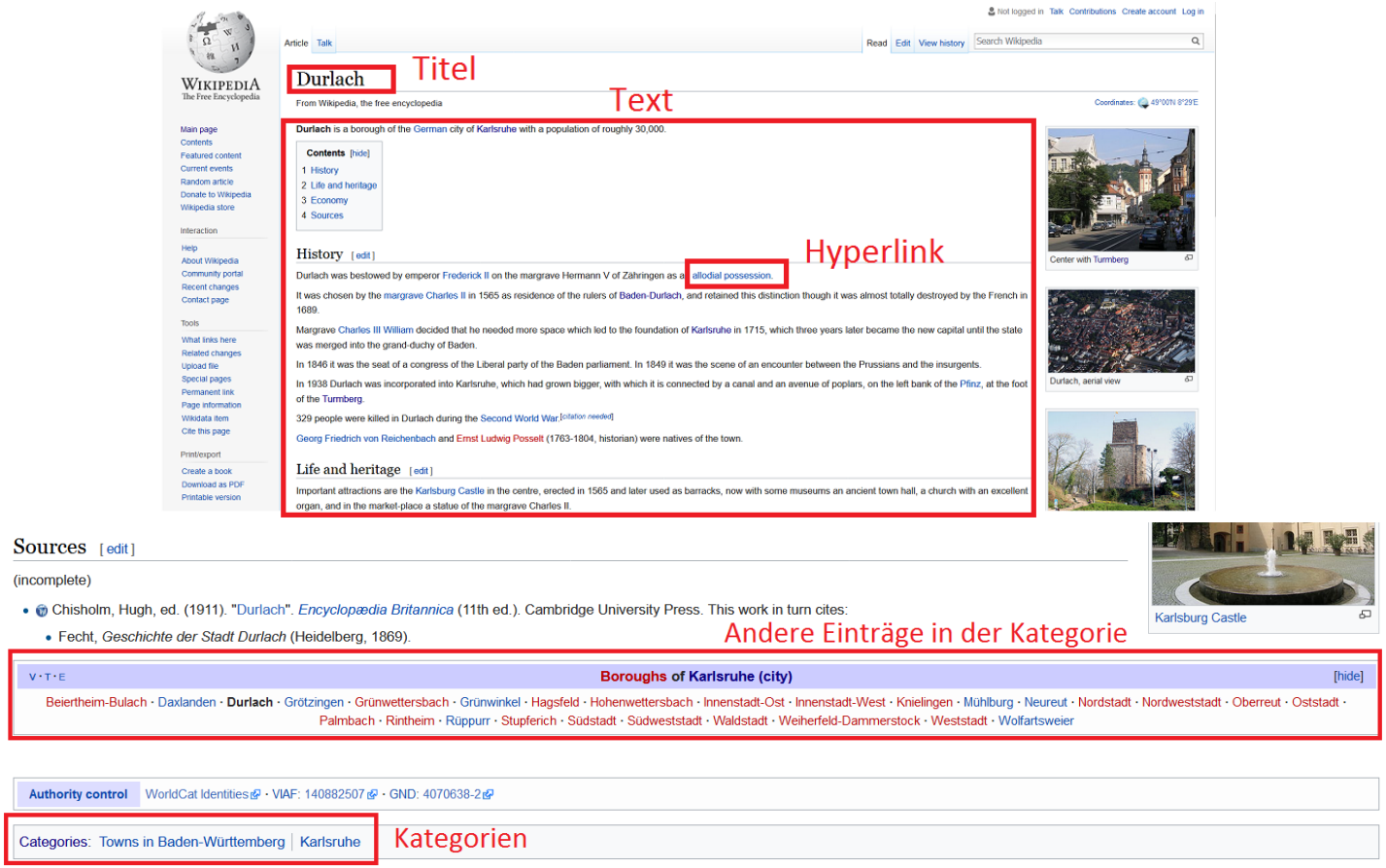
$$Similarity = 2 * \frac{depth\left(kleinster\ gemeinsamer\ Oberbegriff\right)}{depth\left(Konzept\ 1\right) + depth\left(Konzept\ 2\right)} \quad (1)$$

Befragung der externen Datenquellen

- Für jede zwei semantische Ähnliche Domänenwerte (z.B. **Berlin** und **Las Vegas**)
  - Kleinsten Oberbegriff ermitteln (**City**)
- Für jeden Wert in der Liste der Oberbegriffe (z.B. **City**):
  - Seine Kinder in die Liste der Domänenwerte aufnehmen (**Frankfurt**, **Moscow**, **Sydney** usw.)
  - Falls dieser Wert im ersten Schritt ermittelt wurde, werden seine Kinder bis zur Tiefe **n** ausgegeben.
  - **n** - Maximaler Abstand zwischen dem Oberbegriff und Domänenwerten

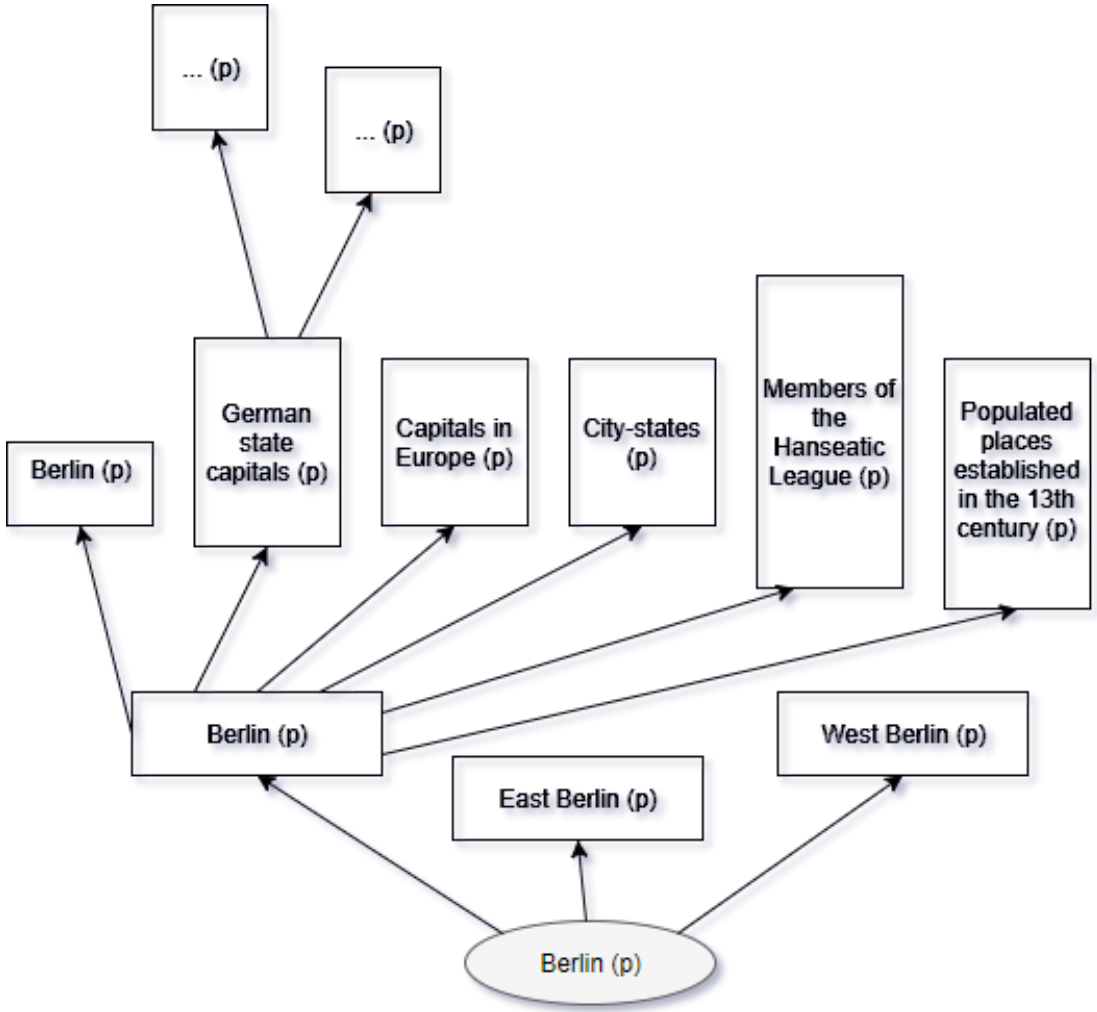
Befragung der externen Datenquellen

Wikipedia:



Befragung der externen Datenquellen

Wikipedia:



Befragung der externen Datenquellen

- Gefundener Artikel: “Moscow (city)”
- “(city)” ist ein bekannten Domänenwert
- “Moscow” aufnehmen und die Bewertung von Moscow erhöhen

Befragung der externen Datenquellen

ResearchCyc:

- Ontologie des Allgemeinwissens
- Besteht aus Konzepten (Cyc-Konstanten)
- Regeln
  - Student ist ein Mensch
- Inferenzmaschine
  - Fakt 1: Student ist ein Mensch
  - Fakt 2: Michael ist Student
  - Schlussfolgerung: Michael ist ein Mensch

Befragung der externen Datenquellen

Befragung von ResearchCyc:

- Für jeden möglichen Oberbegriff (z.B. **City**)
  - Abfragen, welche Konzepte dieser Oberbegriff sein können
  - Beispiel: Was kann ein City sein? Antwort: Berlin ist ein City, Moskau ist ein City usw.
- Für die bekannten Domänenwerte (z.B. **Frankfurt** und **Moskau**)
  - Abfragen, was sowohl **Frankfurt** als auch **Moskau** sind.
  - Antwort: Stadt, Megalopolis, Ort, geographische Entität.
  - Abfragen, welche Konzepte allen ermittelten Gruppen gehören:
  - In diesem Fall: Was noch gleichzeitig eine Stadt, Megalopolis, Ort und geographische Entität sind?
  - Antwort: **New York**, **Sydney**, **Philadelphia** usw.

Konsolidierung der Ergebnisse

- Zusammenführung der Domänenwerte in eine Liste
- Wenn ein Domänenwert auf einen bekannten Oberbegriff endet, Bewertung erhöhen

Evaluation

- Eingabe: Attributewerte der Textfelder
- Die ersten 10, 50, 100 Ausgabewerte betrachtet
- Jeder Wert manuell überprüft
- Testfall 1: Fluckticketanbieter
  - Gesucht: Menge der Städte, Länder, Flughafennamen und -Kodierungen
- Testfall 2: Bahnticketanbieter
  - Gesucht: Menge der Bahnhaltestellen
- Testfall 3: Hotelbuchung
  - Gesucht: Menge der Städte, Hotelnamen und Örtlichkeiten

Evaluation

Tabelle: Trefferquoten bei Testfall 1, Flugticketanbieter

Verwendete Datenquelle	Trefferquote (Top 10)	Trefferquote (Top 50)	Trefferquote (Top 100)
WordNet	0.8	0.76	0.73
Wikipedia	0.8	0.76	0.49
ResearchCyc	0.8	0.96	0.98
Alle Datenquellen	0.9	0.84	0.92

Evaluation

Tabelle: Trefferquoten bei Testfall 2, Bahnhaltestellen

Verwendete Datenquelle	Trefferquote (Top 10)	Trefferquote (Top 50)	Trefferquote (Top 100)
WordNet	0	0	0
Wikipedia	1.0	1.0	1.0
ResearchCyc	0	0	0
Alle Datenquellen	0	0	0

Evaluation

Tabelle: Trefferquoten bei Testfall 3, Hotels

Verwendete Datenquelle	Trefferquote (Top 10)	Trefferquote (Top 50)	Trefferquote (Top 100)
WordNet	0.5	0.52	0.41
Wikipedia	0.6	0.3	0.26
ResearchCyc	10	1.0	0.84
Alle Datenquellen	0.6	0.64	0.62

Zusammenfassung

- Aufgabe der Arbeit:
  - Ermittlung der Wertebreiche von Textfeldern mithilfe der externen Datenquellen
- Lösung:
  - Klassifizierung der Attributwerte auf Oberbegriffe und Domänenwerte
  - Lexikalische Filterung
  - Ermittlung weitere Oberbegriffe der bekannten Domänenwerte
  - Ermittlung weitere Domänenwerte
- Datenquellen:
  - ResearchCyc
  - Wikipedia
  - WordNet
- Die entwickelte Lösung:
  - Liefert eine große Anzahl der Ergebnisse
  - Geeignet für mehrere Anwendungsdomänen

Ausblicke

- Weitere Datenquellen sind zu überlegen
  - YAGO, Wikidata, Google Knowledge Graph
  - Listen in Wikipedia (z.B. “List of cities”)
- Mehr Stichproben zur Verbesserung der Konfigurationsparameter

References I

Geetha, TV (2017). “Unsupervised Domain Ontology Learning from Text”. In: *Mining Intelligence and Knowledge Exploration: 4th International Conference, MIKE 2016, Mexico City, Mexico, November 13-19, 2016, Revised Selected Papers*. Bd. 10089. Springer, S. 132.