Karlsruher Institut für Technologie

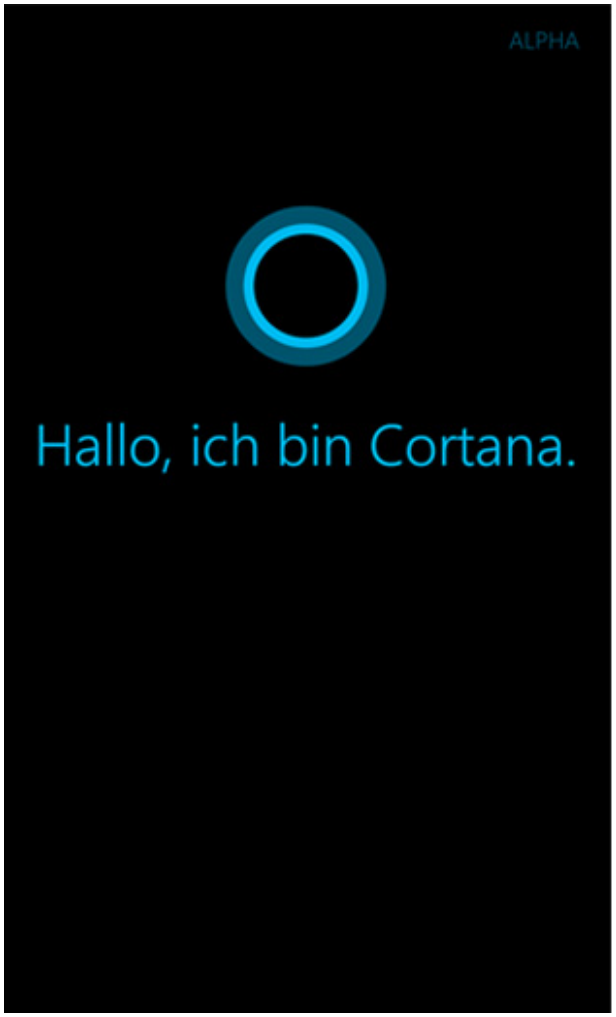
Bachelorarbeit

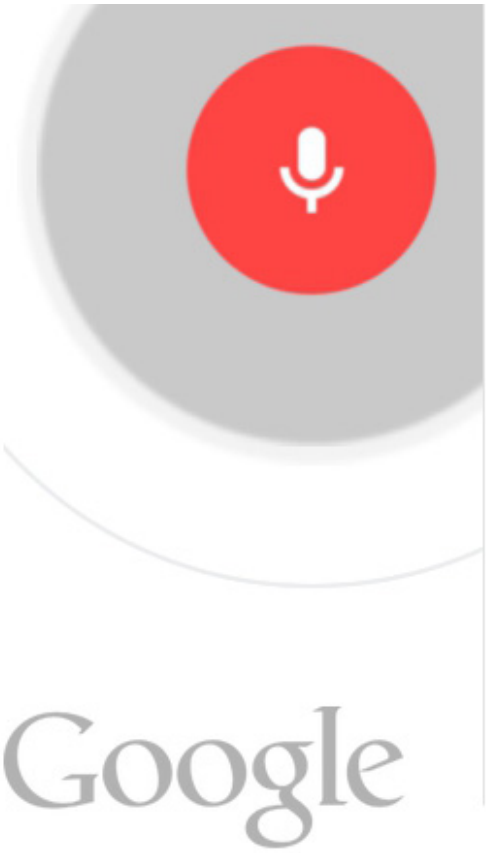
Extraktion und Konsolidierung von Webformularen zur Erzeugung von aktiven Ontologien

Thomas Mayer
Betreut von Martin Blersch

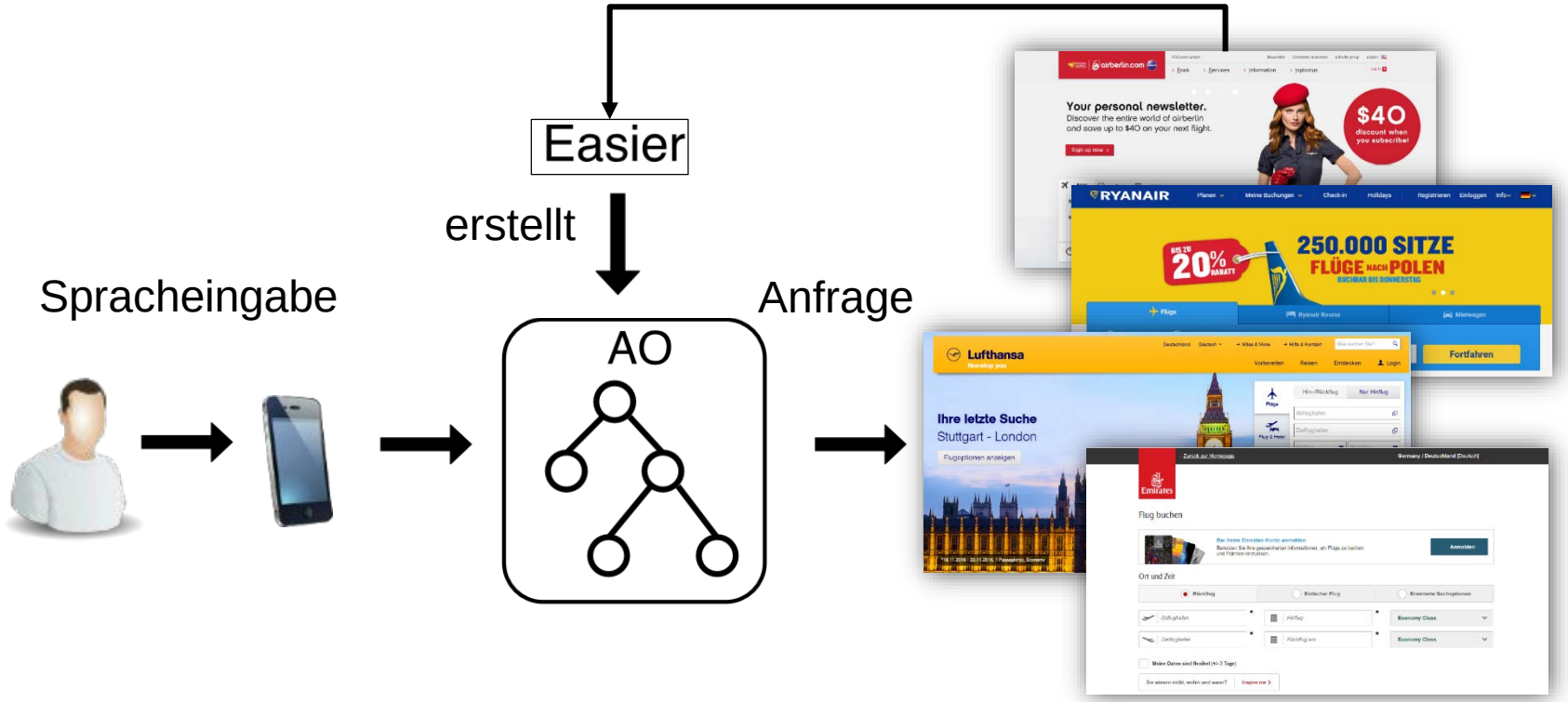
IPD Tichy, Fakultät für Informatik

Sprachassistenten



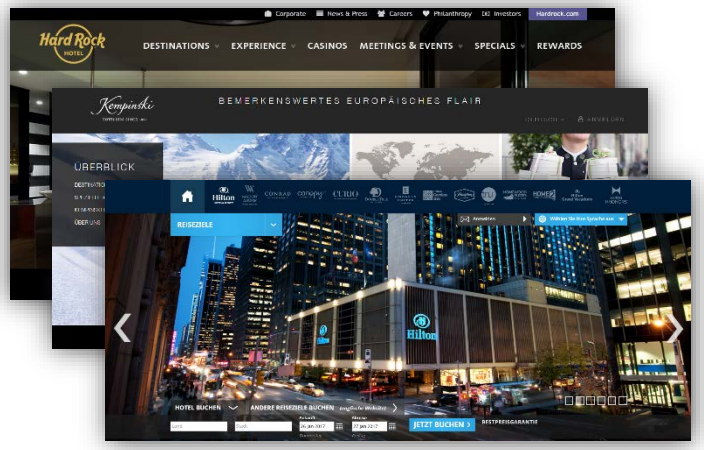


Motivation

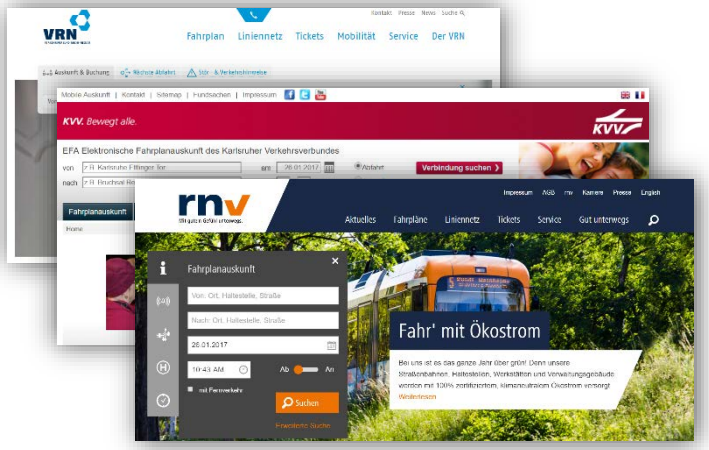

Easier
erstellt
Spracheingabe
Anfrage
AO
Your personal newsletter.
Discover the entire world of airberlin and save up to $40 on your next flight.
$40
discount when you subscribe!
RYANAIR
250.000 SITZE
FLÜGE NACH POLEN
Fortfahren
Lufthansa
Ihre letzte Suche
Stuttgart - London
Flugoptionen anzeigen
Flug buchen
Ort und Zeit

Dienstkategorie

Hotel

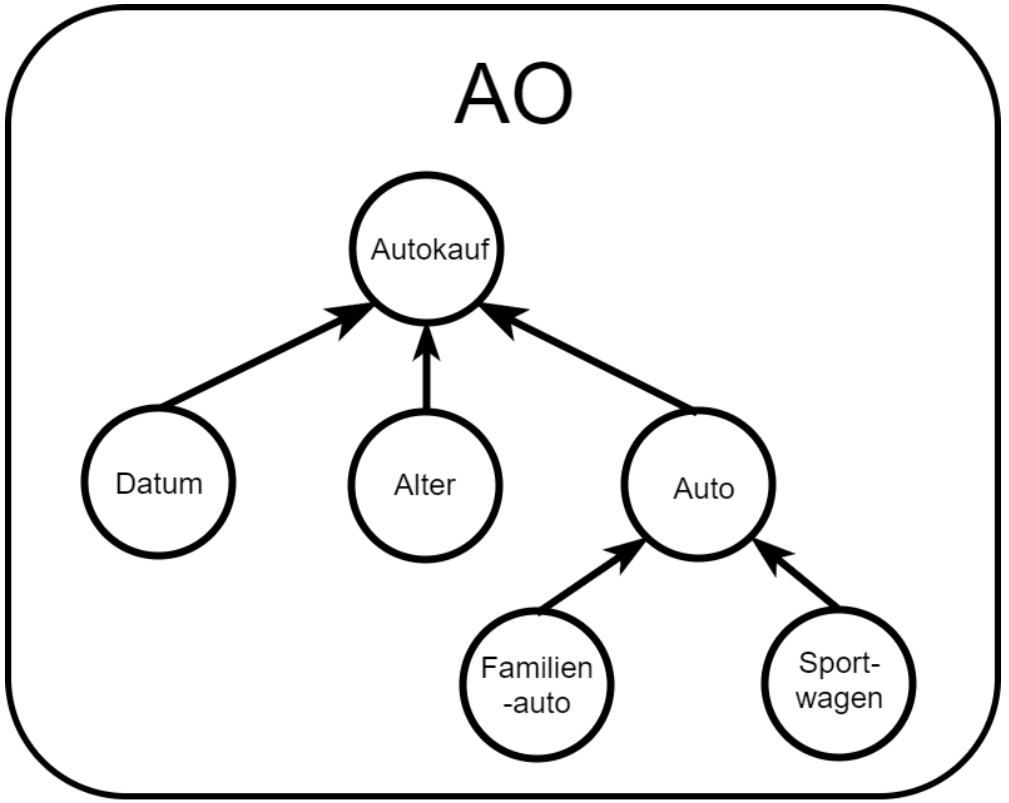


Bahn



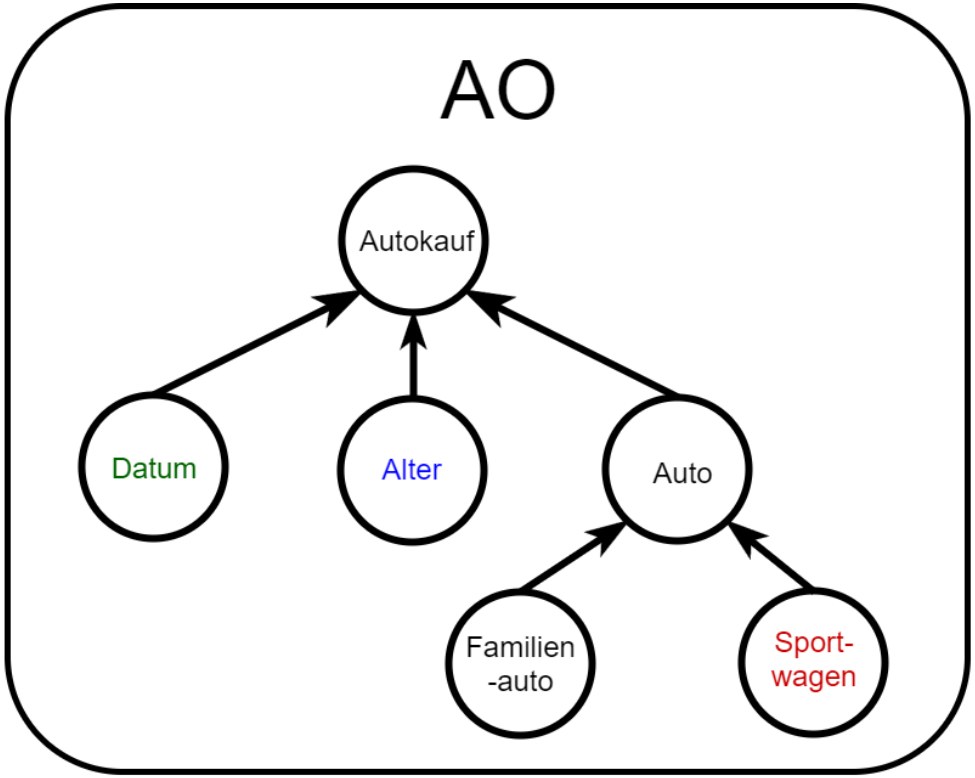
Aktive Ontologien

- Sensorknoten
- Sammel-Knoten
- Auswahl-Knoten



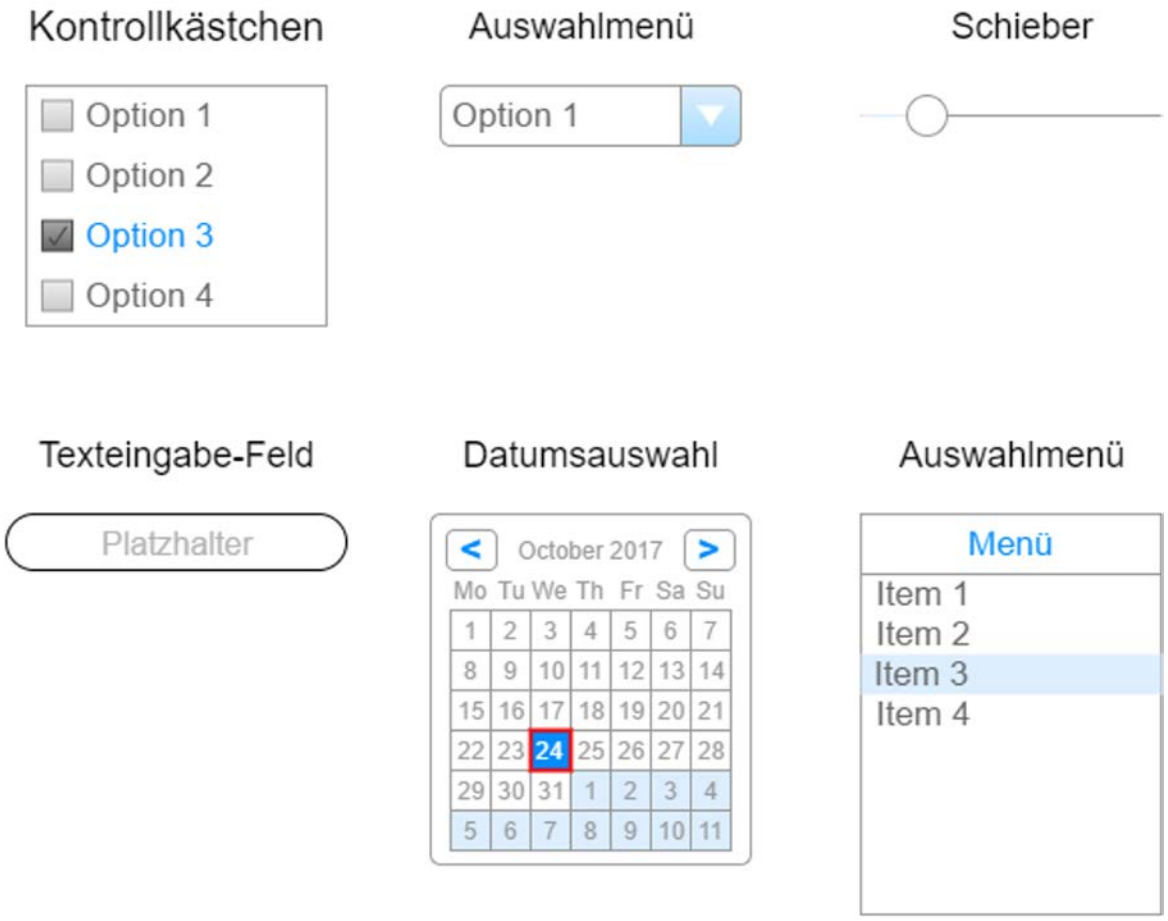
Sensorknoten

- Vokabel-Knoten
- Präfix-Knoten
- Postfix-Knoten
- Knoten mit regulären Ausdrücken
- Spezialisierte Knoten



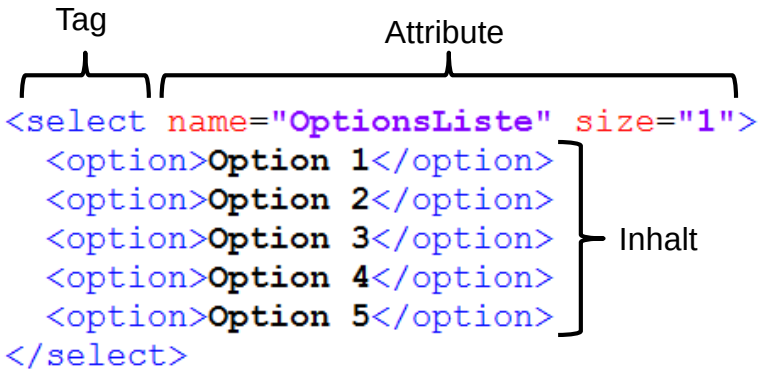
Ich will am Montag einen 20 Jahre alten Ferrari kaufen.

Formularelemente

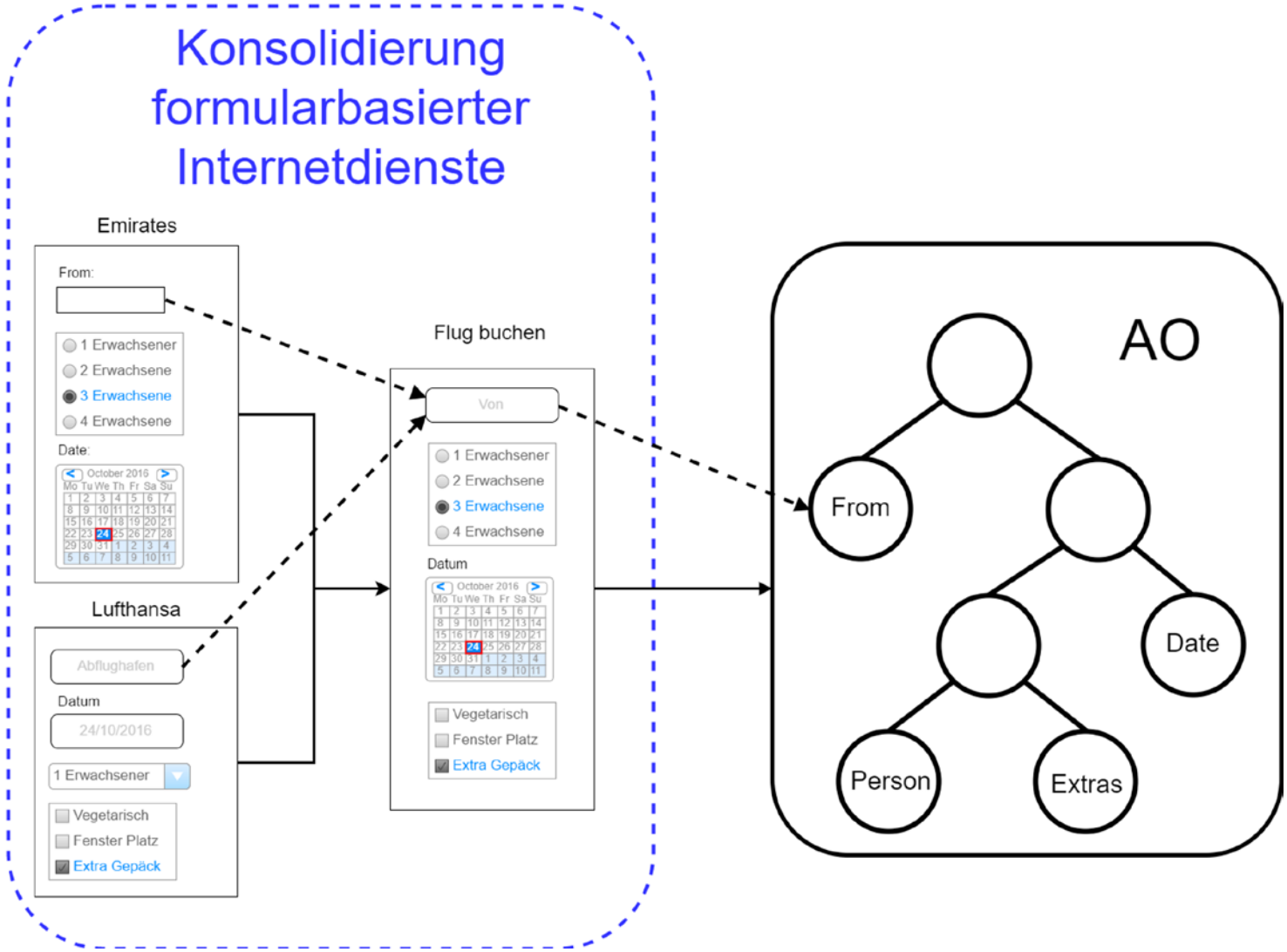


Bestandteile eines Formularelementes

- Tag
  - Auswahlliste, Texteingabefeld, Taste,…
- Attribute
  - Name, Platzhalter, Größe, Darstellung,…
- Inhalt
  - Beschreibung des Elementes, eine Liste von Optionen,…



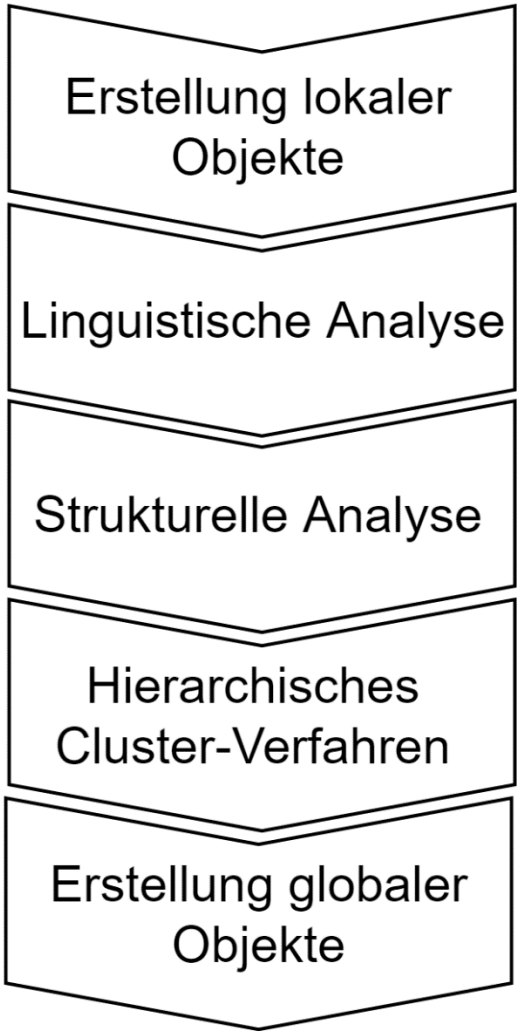
Ziel dieser Arbeit



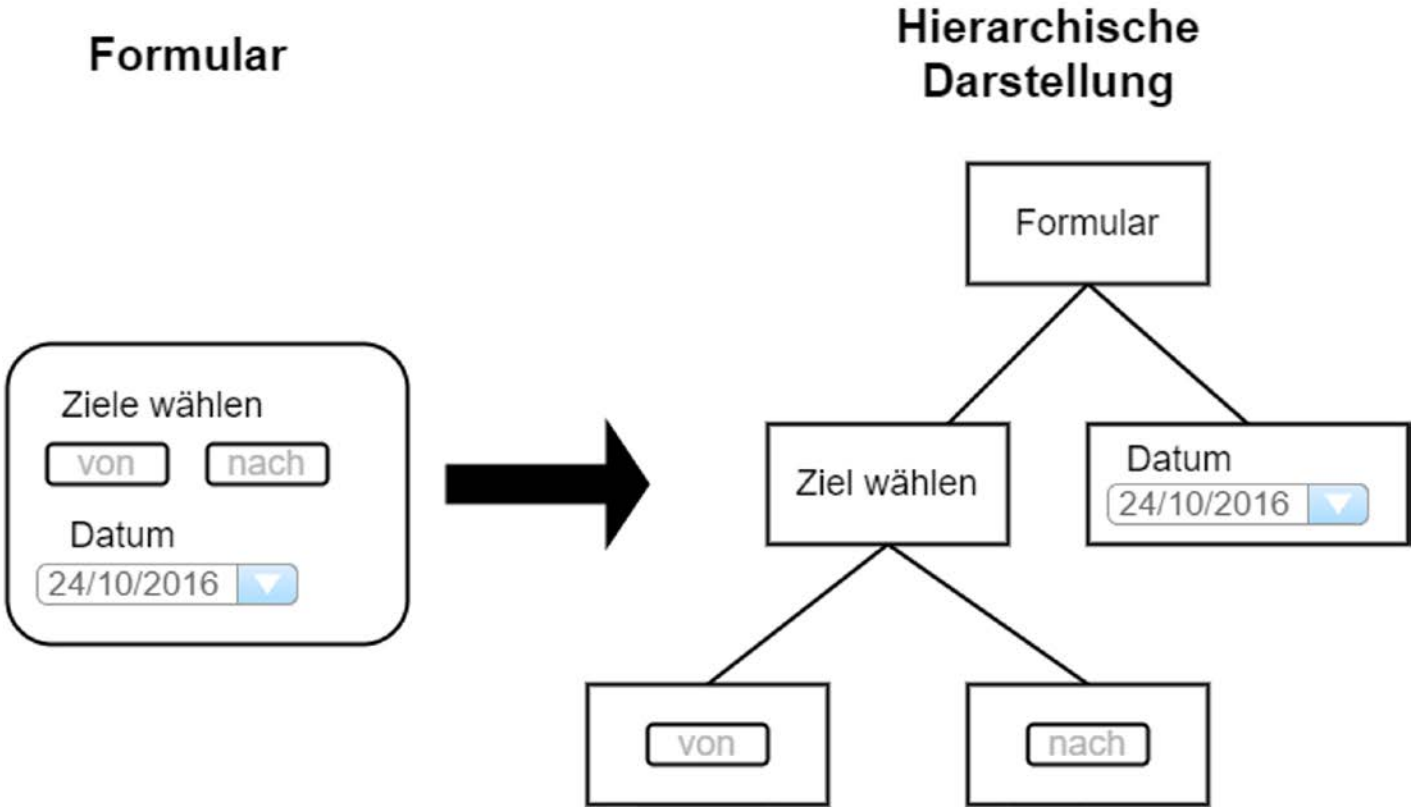
Verwandte Arbeiten

- Konsolidierung von HTML-Formularen im Bereich des Deep Webs
  - WISE-Integrator [WYDM04] [HHYW05]
- Ontology Matching
  - [AG05]
- Schema Matching
  - Cupid [JM01]

Lösungsansatz



Erstellung lokaler Objekte



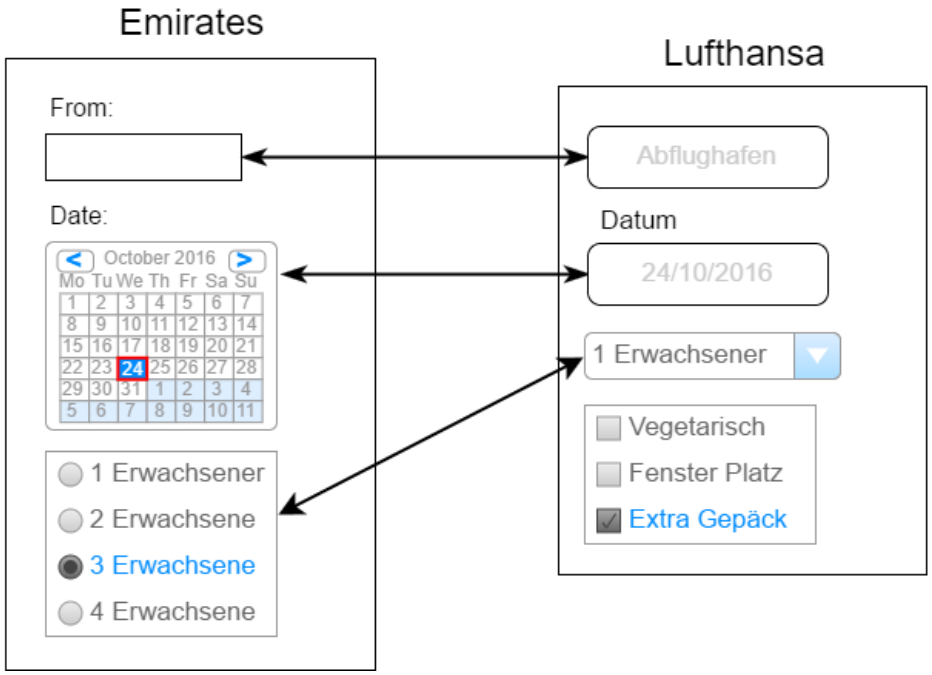
Linguistische Analyse

- Bestandteile
  - Token-Analyse
    - Namensgebende und beschreibende Attribute
    - Vergleicht Wörter
  - Teilzeichenketten-Analyse
    - Namensgebende und beschreibende Attribute
    - Sucht größte Teilzeichenkette
  - Werte-Analyse
    - Vergleicht die Eingabewerte

```
<textarea rows="4" cols="50" name="TextFeld">
</textarea>
```

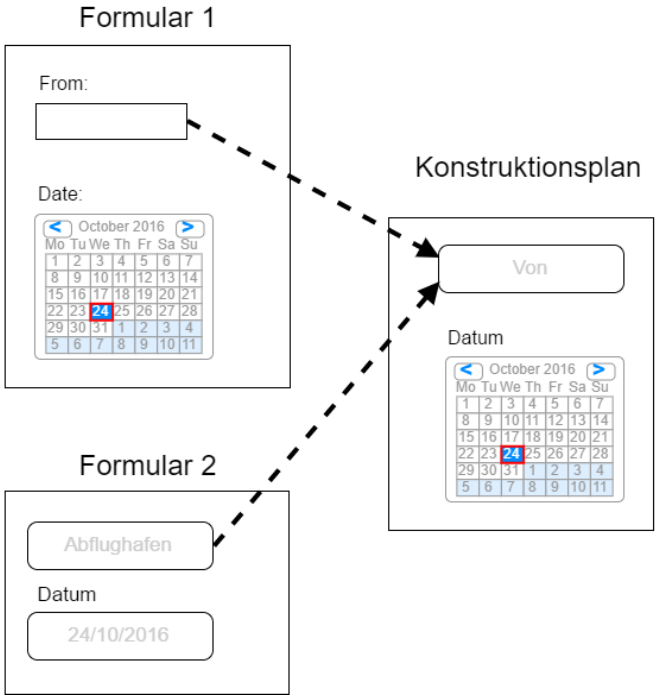
Strukturelle Analyse

- Erhält die Ähnlichkeitswerte der linguistischen Analyse als Eingabe
  - Idee: Bestimmung der Ähnlichkeiten der Formularelemente anhand ihrer Position innerhalb der Formulare
    - Die strukturelle Ähnlichkeit eines Formularelementes wird anhand der linguistischen Ähnlichkeiten der Nachbarelemente gemessen

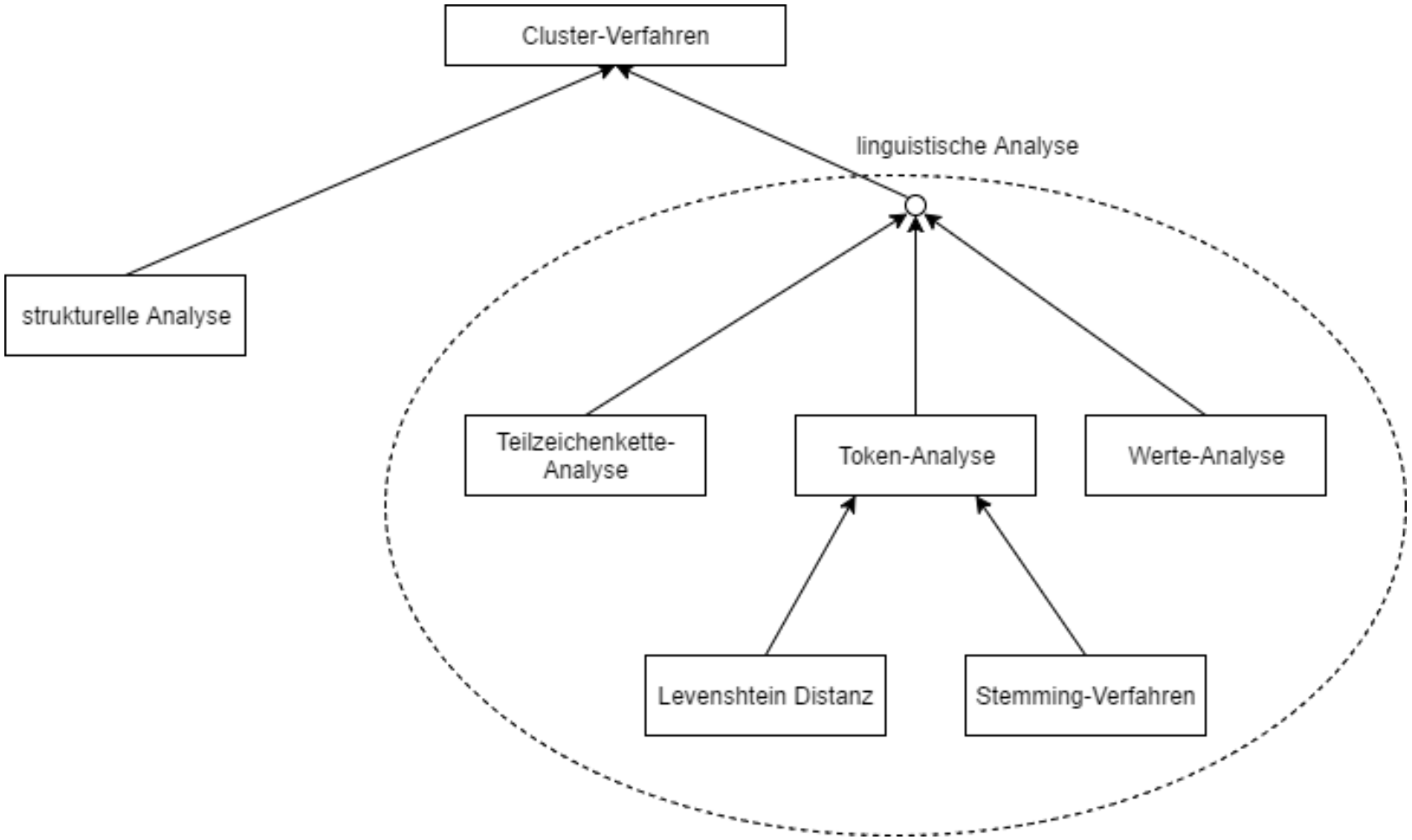


Hierarchisches Cluster-Verfahren und Erstellung globaler Objekte

- Mithilfe der Ähnlichkeitswerte werden Cluster erstellt, welche semantisch gleiche Formularelemente enthalten
- Für jedes Cluster wird ein globales Objekt erstellt, welches zu dem Konstruktionsplan hinzugefügt wird



Verschiedene Verfahren

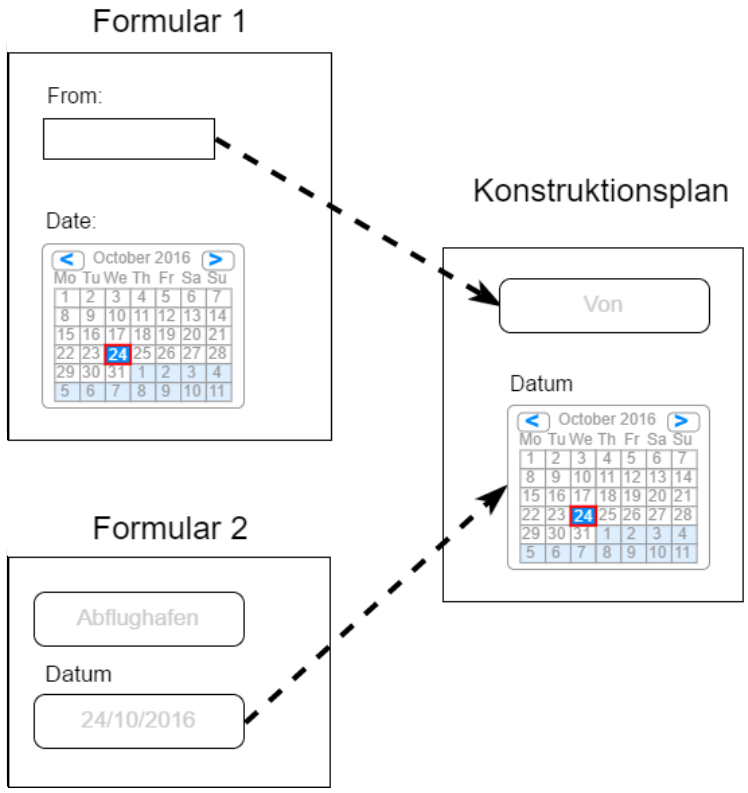


Aufbau Evaluation

- Finden geeigneter Parameterwerte
- Trainingsmenge
  - Kategorie Flug (3 Formulare)
- Erstellung eines Goldstandards
- Erstellung von Testmengen
  - Kategorie Bahn und Kategorie Hotel (jeweils 10 Formulare)

Bewertung

- Richtig positiv
  - Eine richtige Abbildung
- Falsch positiv
  - Eine falsche Abbildung
- Falsch negativ
  - Eine nicht abgebildete Abbildung
- Richtig negativ
  - Eine nicht vorhandene Abbildung



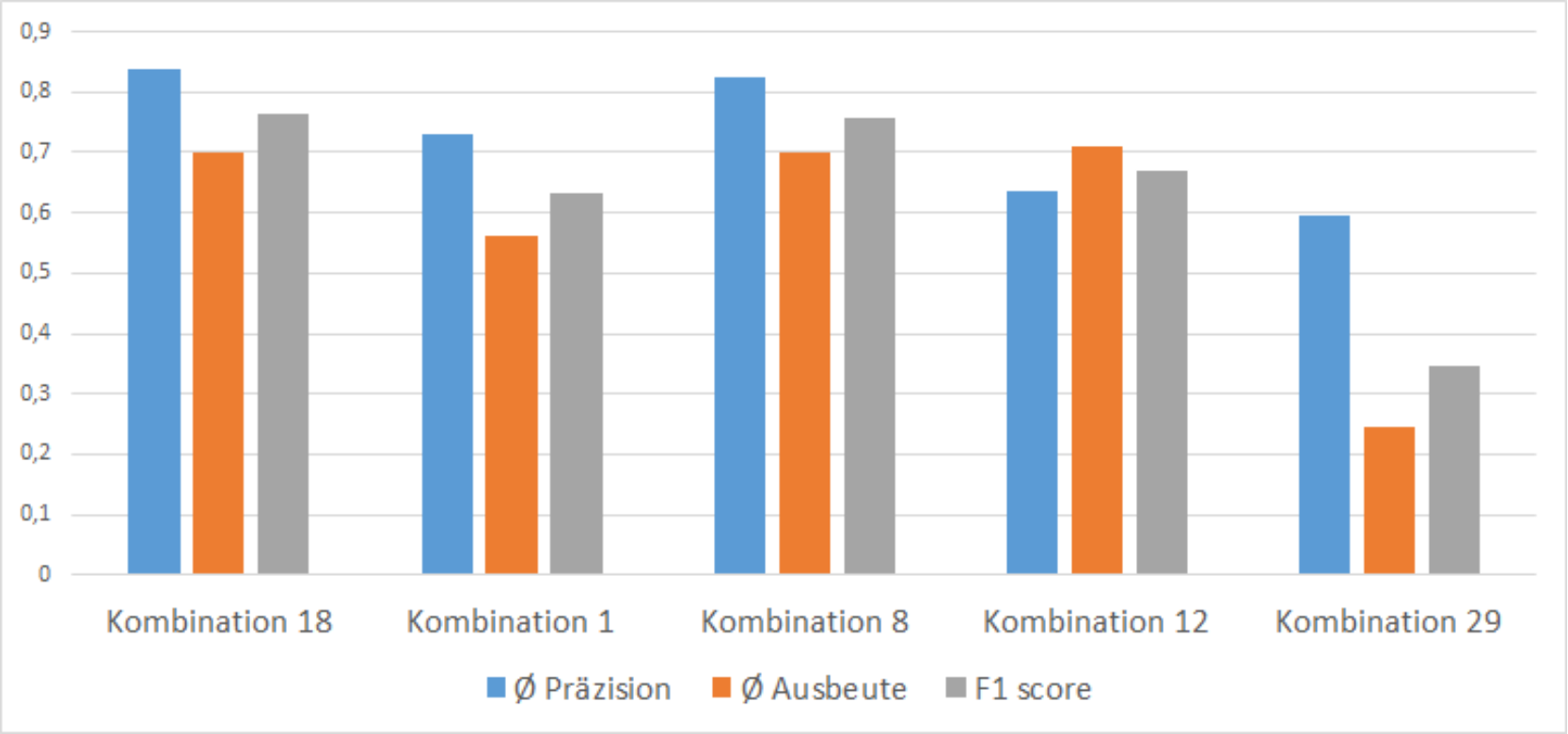
Ergebnisse der Testmengen

*

	Präzision	Ausbeute	F1 score
Bahn	94,1%	76,2%	84,4%
Hotel	73,7%	63,6%	68,3%
Ø	83,9%	69,9%	76,3%

* Verwendet wurde Kombination 18 (Teilzeichenketten-Analyse und strukturelle Analyse)

Ergebnisse der Testmengen



- 18 Teilzeichenketten-Analyse, strukturelle Analyse
- 1 Alle Verfahren
- 8 Token-Analyse, Teilzeichenketten-Analyse, strukturelle Analyse
- 12 Token-Analyse, Werte-Analyse, strukturelle Analyse
- 29 Token-Analyse, Levenshtein Distanz, Werte-Analyse

Fazit

- Ziel dieser Arbeit
  - Automatische Konsolidierung formularbarsierter Internetdienste
  - Automatische Erstellung eines Konstruktionsplans
- Lösung
  - Linguistische und strukturelle Analyse
  - Cluster-Verfahren
- Ergebnisse
  - Ausbeute 70%
  - Präzision 84%

Ausblick

- Verwendung von Wörterbüchern
- Verwendung von Synonym-Tabellen
- Erstellung von komplexen Abbildungen

Literatur

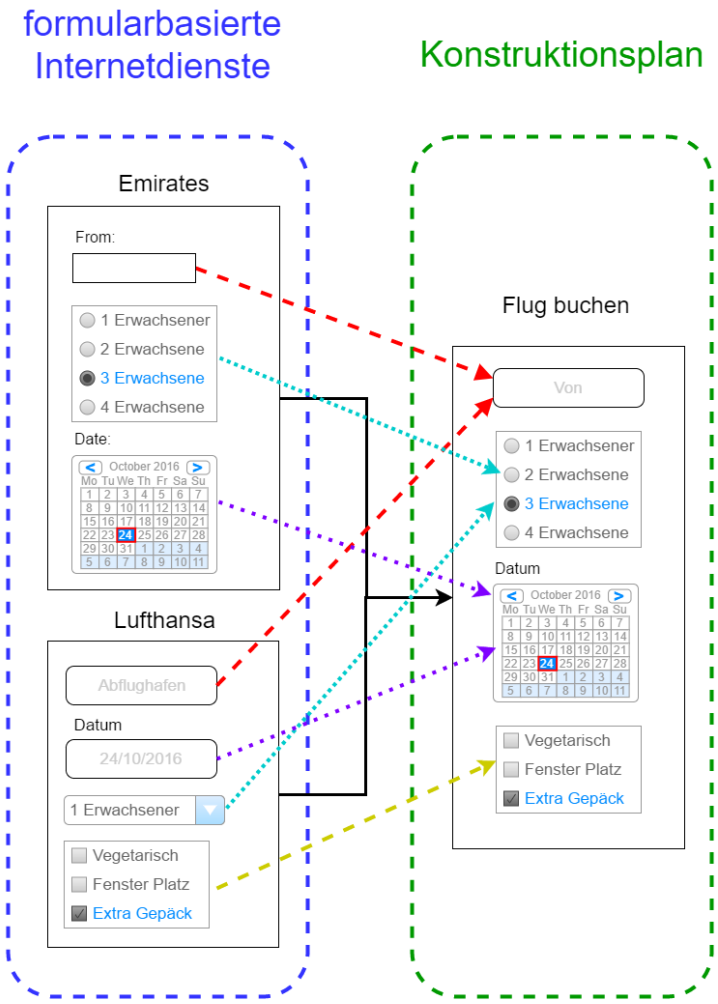
- [AG04] Avigdor Gal, Hasan J. Giovanni Modica M. Giovanni Modica: OntoBuilder: Fully Automatic Extraction and Consolidation of Ontologies from Web Sources. ICDE Conference : IEEE, 2004.
- [AG05] Avigdor Gal, Hasan Jamil Ami E. Giovanni Modica M. Giovanni Modica: Automatic Ontology Matching Using Application Semantics. In: AI MAGAZINE 26 (2005).
- [BL16] Blersch, M. ; Landhäuser, M.: EASIER: An Approach to Automatically Generate Active Ontologies for Intelligent Assistants. The 20th World Multi-Conference on Systemics, Cybernetics and Informatics (WMSCI 2016) Orlando, FL, USA 05.07.2016 DOI: 10.13140/RG.2.1.2586.9043, Juli 2016.
- [Guz08] Guzzoni, Didier: Active: A Unified Platform for Building intelligent Applications, École polytechnique fédérale de Lausanne, dissertation, 2008.

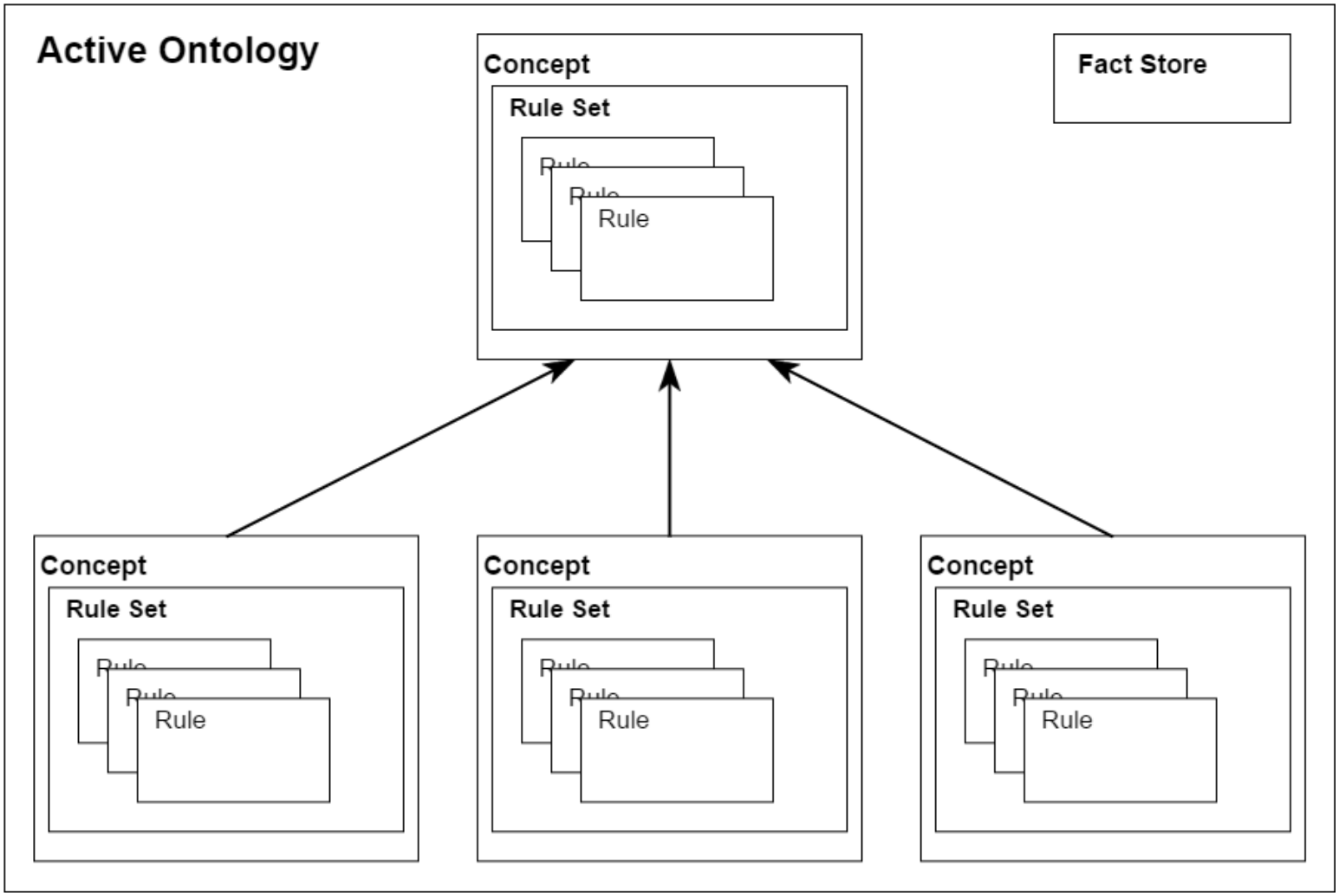
Literatur

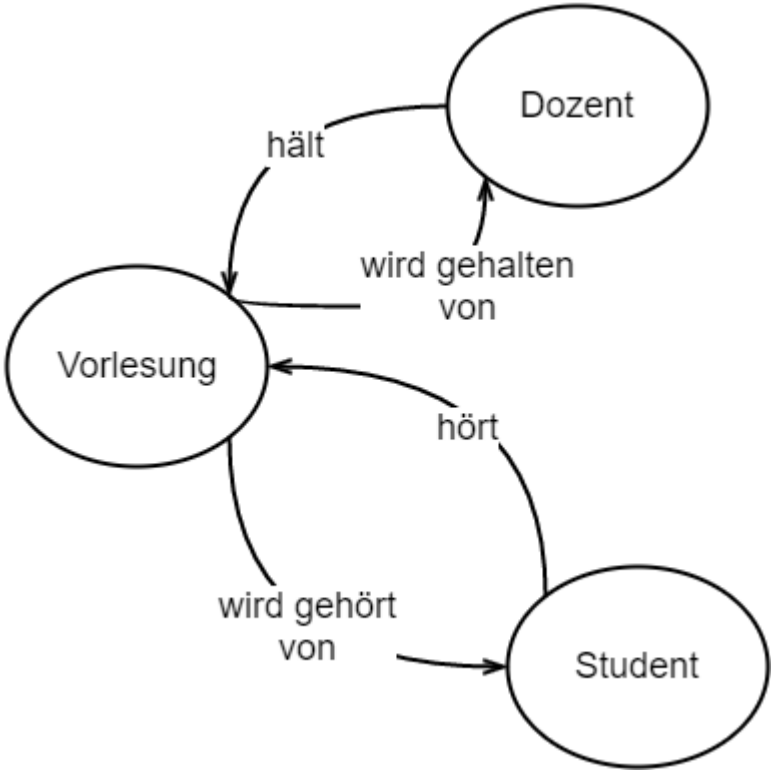
- [HHYW03] Hai He, Weiyi M. (Hrsg.) ; Yu, Clement (Hrsg.) ; Wu, Zonghuan (Hrsg.): WISE-Integrator: An Automatic Integrator of Web Search Interfaces for E-Commerce. Bd. 29. VLDB Endowment, 09 2003.
- [HHYW05] Hai He, Weiyi M. (Hrsg.) ; Yu, Clement (Hrsg.) ; Wu, Zonghuan (Hrsg.): WISE-Integrator: A System for Extracting and Integrating Complex Web Search Interfaces of the Deep Web. VLDB Endowment, 08 2005.
- [JM01] Jayant Madhavan, Erhard R. Phil Bernstein B. Phil Bernstein: Generic Schema Matching With Cupid / Microsoft Research. 2001.
- [MFBB10] Maiz, Nora (Hrsg.) ; Fahad, Muhammad (Hrsg.) ; Boussaid, Omar (Hrsg.); Bentayeb, Fadila (Hrsg.): Automatic Ontology Merging by Hierarchical Clustering and Inference Mechanisms. 2010.

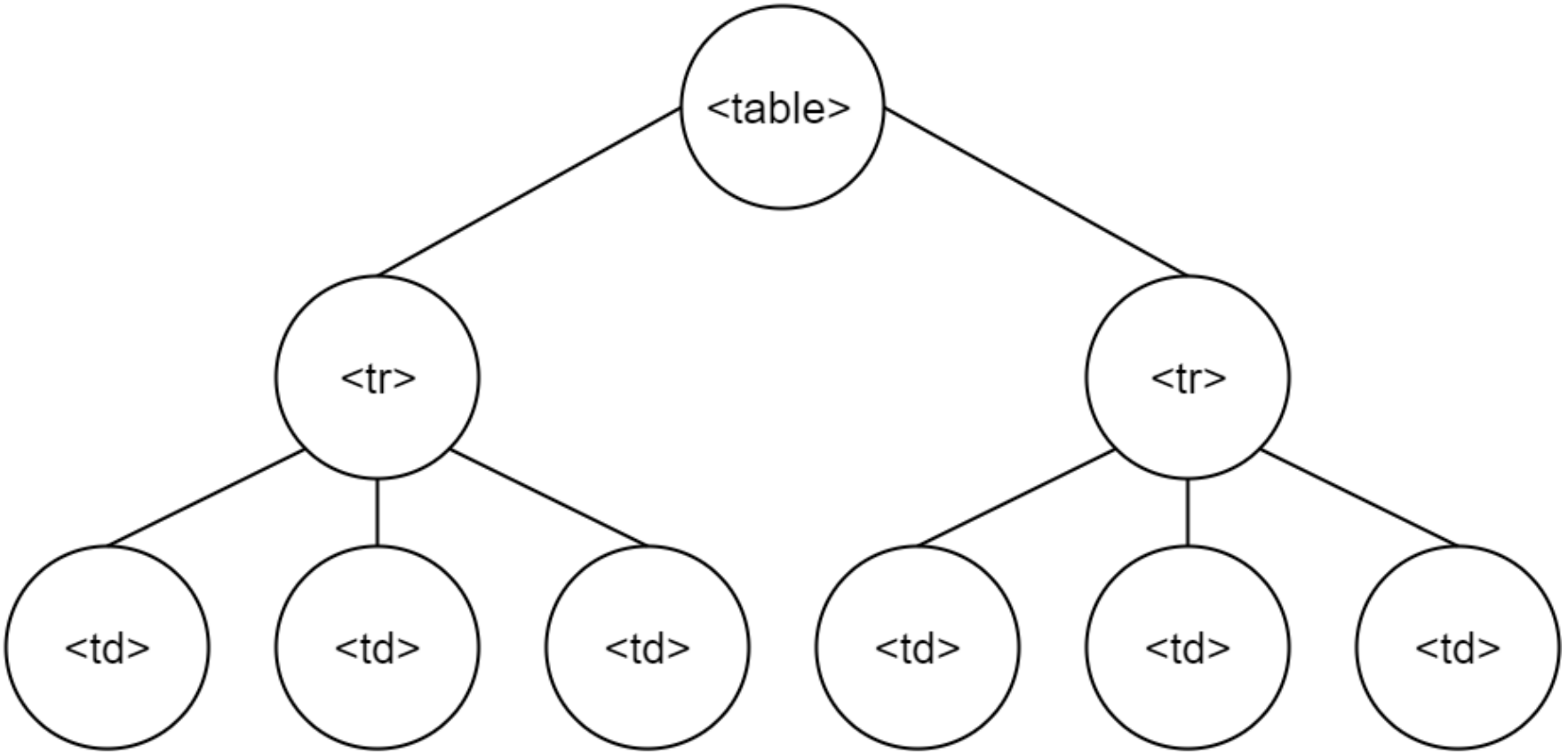
Literatur

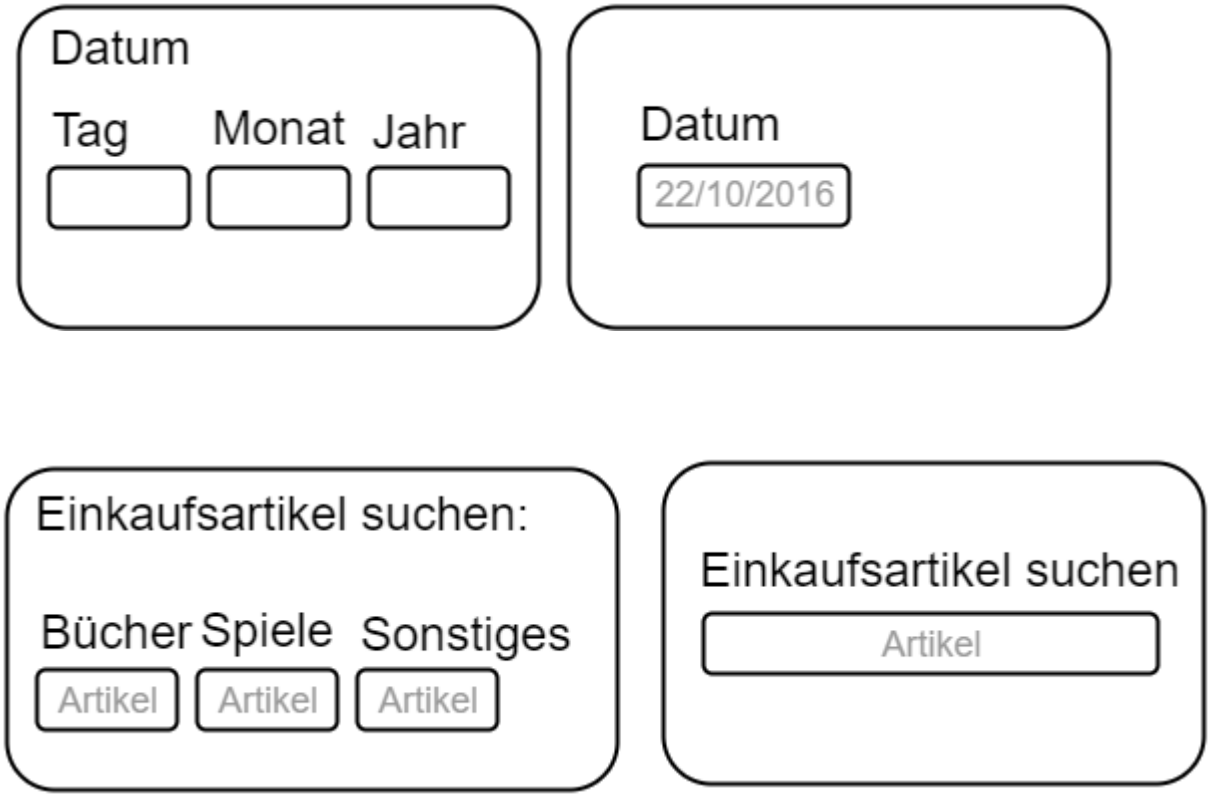
- [Was16] Wasim, Said: Abbildung von Webformularen auf aktive Ontologien. Germany, Karlsruher Institut für Technologie, Lehrstuhl IPD Tichy, Masterarbeit, 2016.
- [Sch16] Schmitteckert, Kay: Semi-automatische Generierung von aktiven Ontologien aus Webformularen. Germany, Karlsruher Institut für Technologie, Lehrstuhl IPD Tichy, Bachelorarbeit, 2016.
- [WYDM04] Wu, Wensheng ; Yu, Clement ; Doan, AnHai ; Meng, Weiyi: An Interactive Clustering-based Approach to Integrating Source Query Interfaces on the Deep Web. In: Proceedings of the 2004 ACM SIGMOD international conference on Management of data Table of Contents, ACM New York, NY, USA 2004, JUL 2004.
- [YA12] Yuan An, Il-Yeol S. Xiaohua Hu H. Xiaohua Hu (Hrsg.): Learning to Discover Complex Mappings from Web Forms to Ontologies. CIKM ’12 Proceedingsof the 21st ACM international conference on Information and knowledge management, 10 2012.

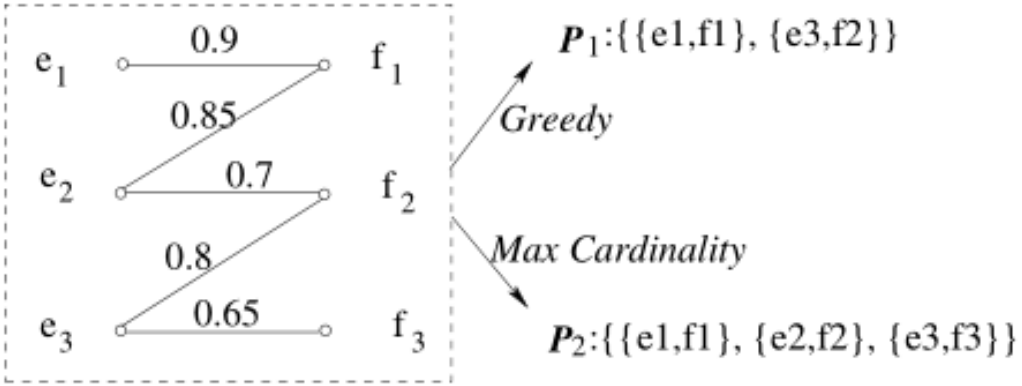

formularbasierte Internetdienste
Konstruktionsplan
Emirates
From:
1 Erwachsener
2 Erwachsene
3 Erwachsene
4 Erwachsene
Date:
October 2016
Lufthansa
Abflughafen
Datum
24/10/2016
1 Erwachsener
Vegetarisch
Fenster Platz
Extra Gepäck
Flug buchen
Von
1 Erwachsener
2 Erwachsene
3 Erwachsene
4 Erwachsene
Datum
October 2016
Vegetarisch
Fenster Platz
Extra Gepäck

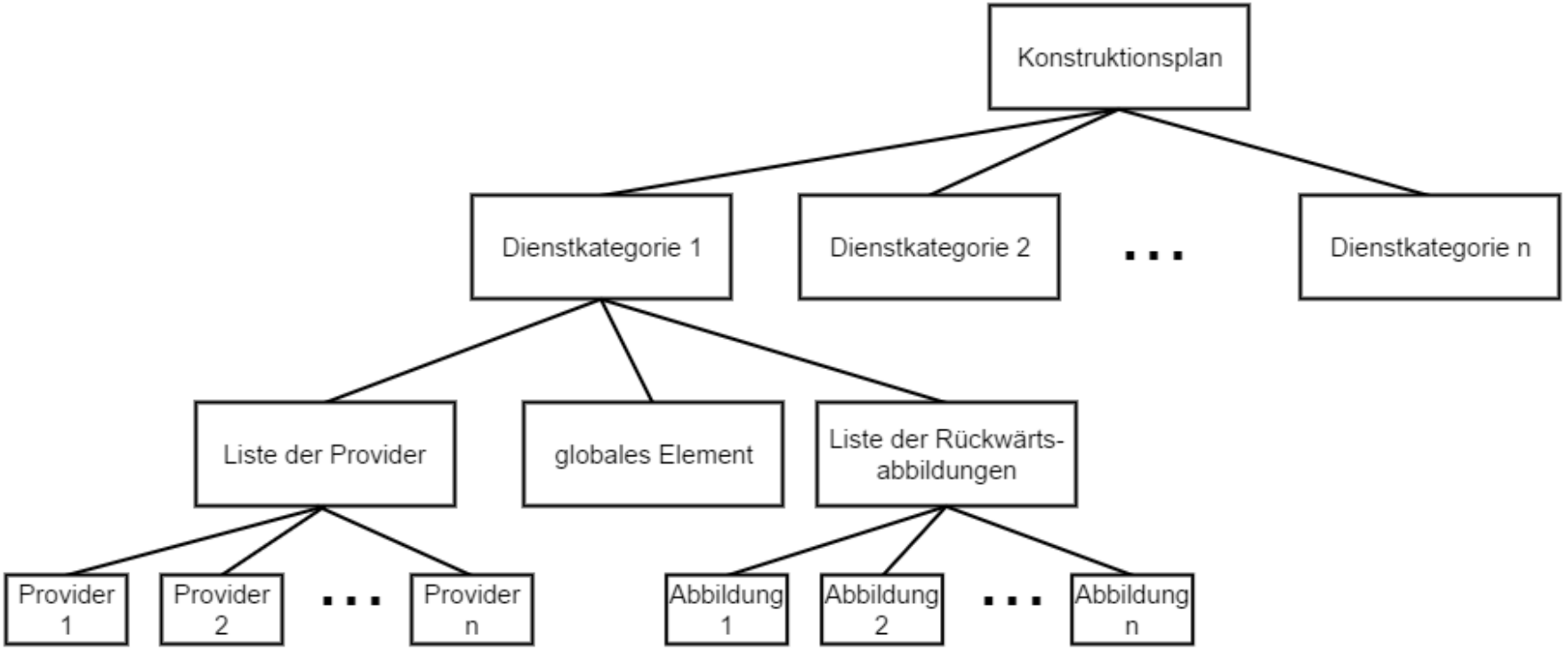

Active Ontology
Concept
Rule Set
Rule
Fact Store
Concept
Rule Set
Rule
Concept
Rule Set
Rule
Concept
Rule Set
Rule

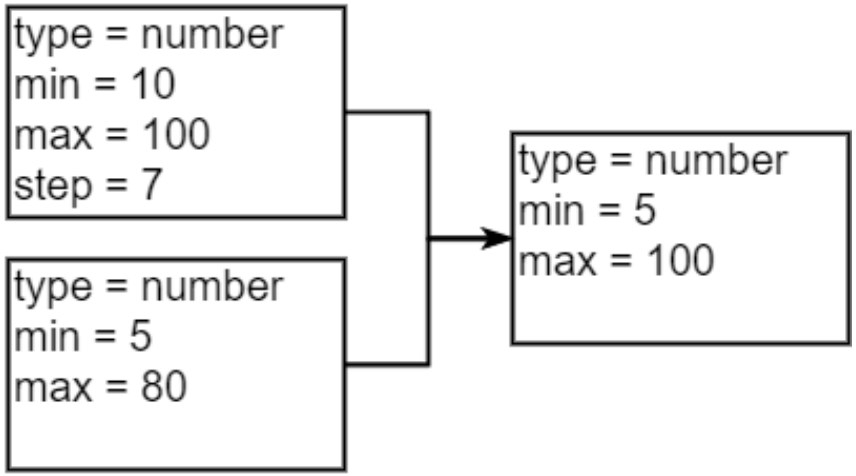

Dozent
hält
wird gehalten
von
Vorlesung
hört
wird gehört
von
Student

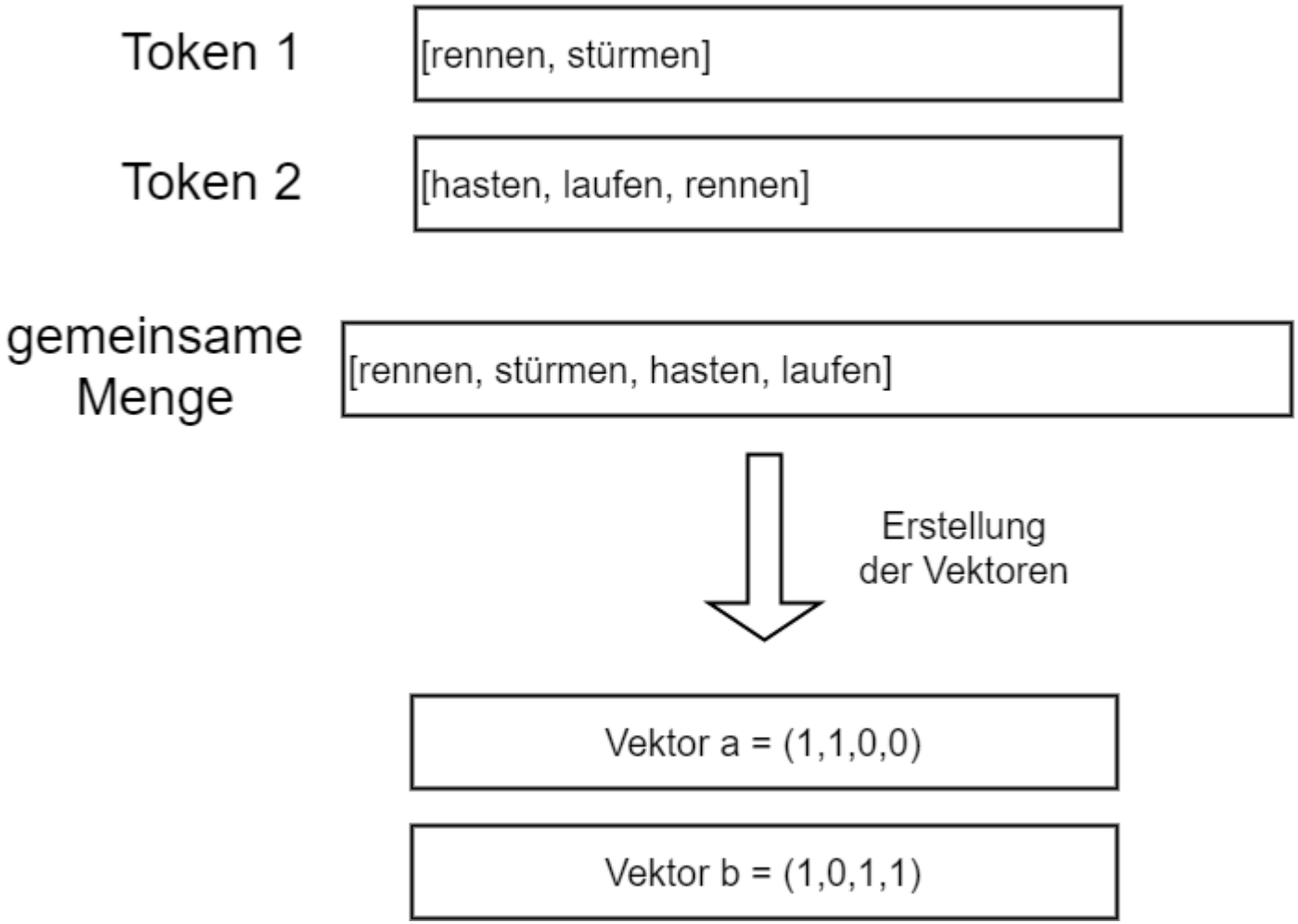

<table>
<tr>
<tr>
<td>
<td>
<td>
<td>
<td>
<td>


Datum
Tag
Monat
Jahr
Datum
22/10/2016
Einkaufsartikel suchen:
Bücher
Spiele
Sonstiges
Artikel
Artikel
Artikel
Einkaufsartikel suchen
Artikel


e1
e2
e3
f1
f2
f3
0.9
0.85
0.7
0.8
0.65
P1:{{e1,f1}, {e3,f2}}
Greedy
Max Cardinality
P2:{{e1,f1}, {e2,f2}, {e3,f3}}


Konstruktionsplan
Dienstkategorie 1
Dienstkategorie 2
. . .
Dienstkategorie n
Liste der Provider
globales Element
Liste der Rückwärts-abbildungen
Provider 1
Provider 2
. . .
Provider n
Abbildung 1
Abbildung 2
. . .
Abbildung n


type = number
min = 10
max = 100
step = 7
type = number
min = 5
max = 80
type = number
min = 5
max = 100


Token 1
[rennen, stürmen]
Token 2
[hasten, laufen, rennen]
gemeinsame Menge
[rennen, stürmen, hasten, laufen]
Erstellung der Vektoren
Vektor a = (1,1,0,0)
Vektor b = (1,0,1,1)

$$Cos(a,b) = \frac{a \bullet b}{||a|| * ||b||}$$

Vektor a = (1,1,0,0)

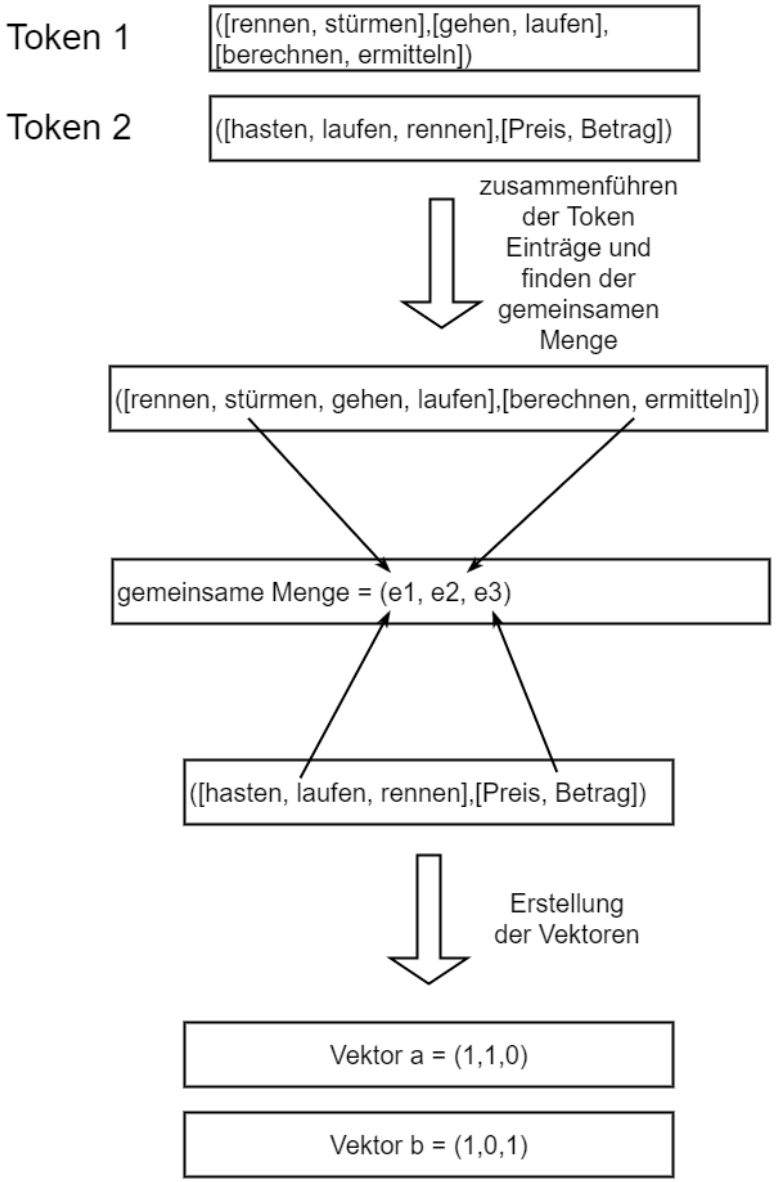
Vektor b = (1,0,1,1)

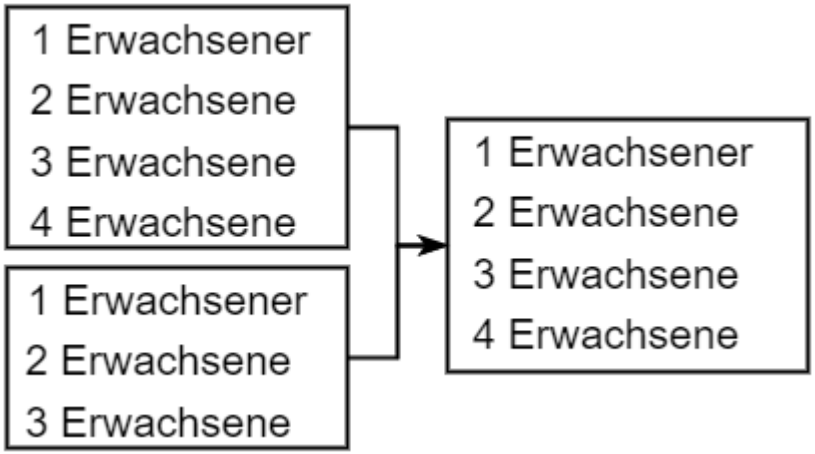
$$a \bullet b = 1 * 1 + 1 * 0 + 0 * 1 + 0 * 1 = 1$$

$$||a|| = \sqrt{1^2 + 1^2 + 0^2 + 0^2} = \sqrt{2}$$

$$||b|| = \sqrt{1^2 + 0^2 + 1^2 + 1^2} = \sqrt{3}$$

$$Cos(a,b) = \frac{a \bullet b}{||a|| * ||b||} = \frac{1}{\sqrt{2} * \sqrt{3}} \approx 0,3178$$

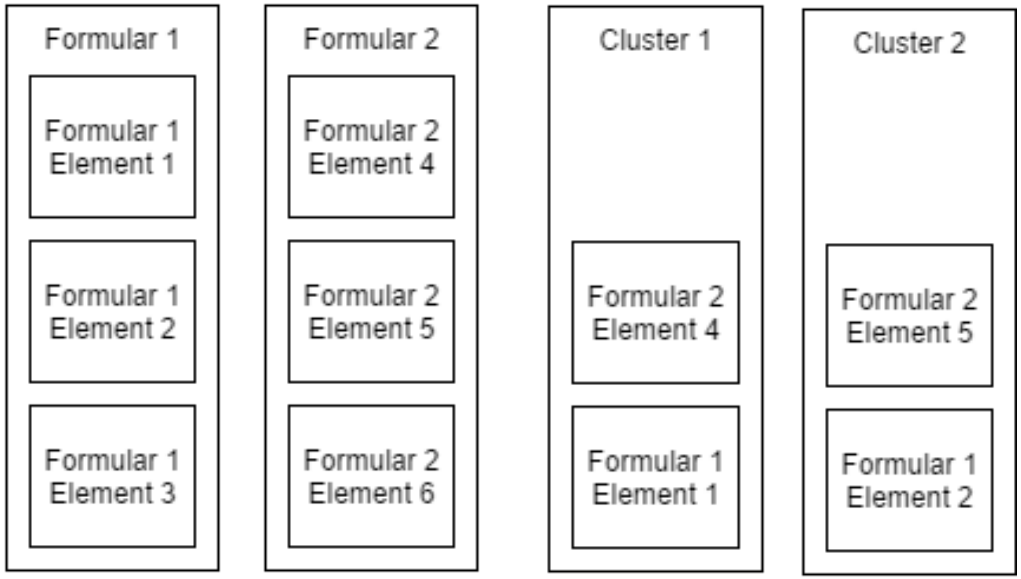

Token 1
([rennen, stürmen],[gehen, laufen], [berechnen, ermitteln])
Token 2
([hasten, laufen, rennen],[Preis, Betrag])
zusammenführen der Token Einträge und finden der gemeinsamen Menge
([rennen, stürmen, gehen, laufen],[berechnen, ermitteln])
gemeinsame Menge = (e1, e2, e3)
([hasten, laufen, rennen],[Preis, Betrag])
Erstellung der Vektoren
Vektor a = (1,1,0)
Vektor b = (1,0,1)

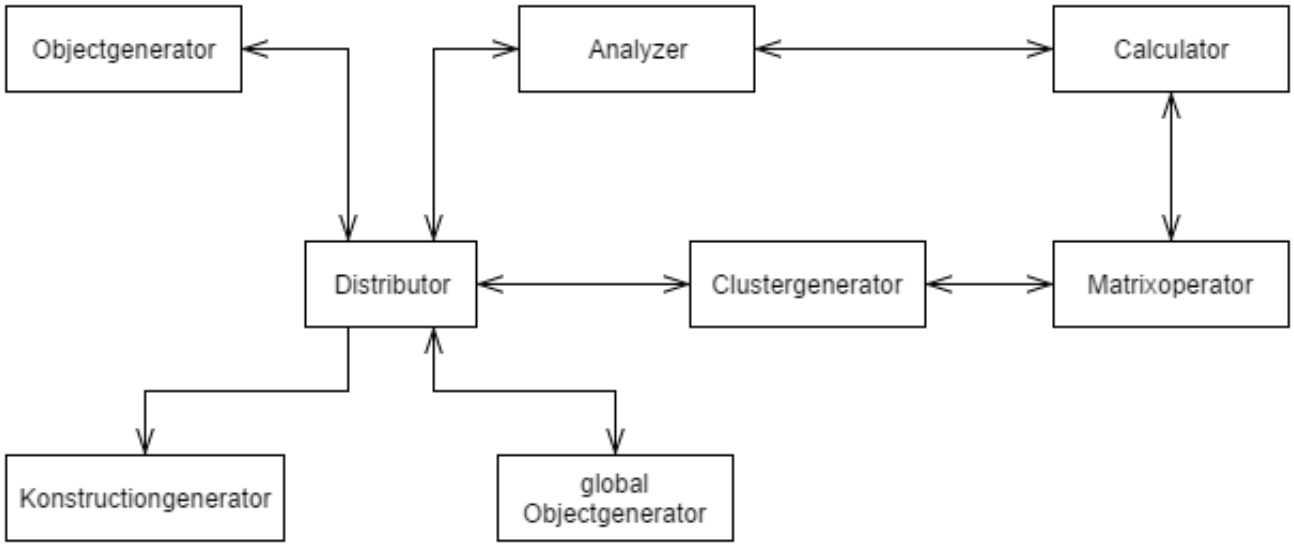

1 Erwachsener
2 Erwachsene
3 Erwachsene
4 Erwachsene
1 Erwachsener
2 Erwachsene
3 Erwachsene
1 Erwachsener
2 Erwachsene
3 Erwachsene
4 Erwachsene

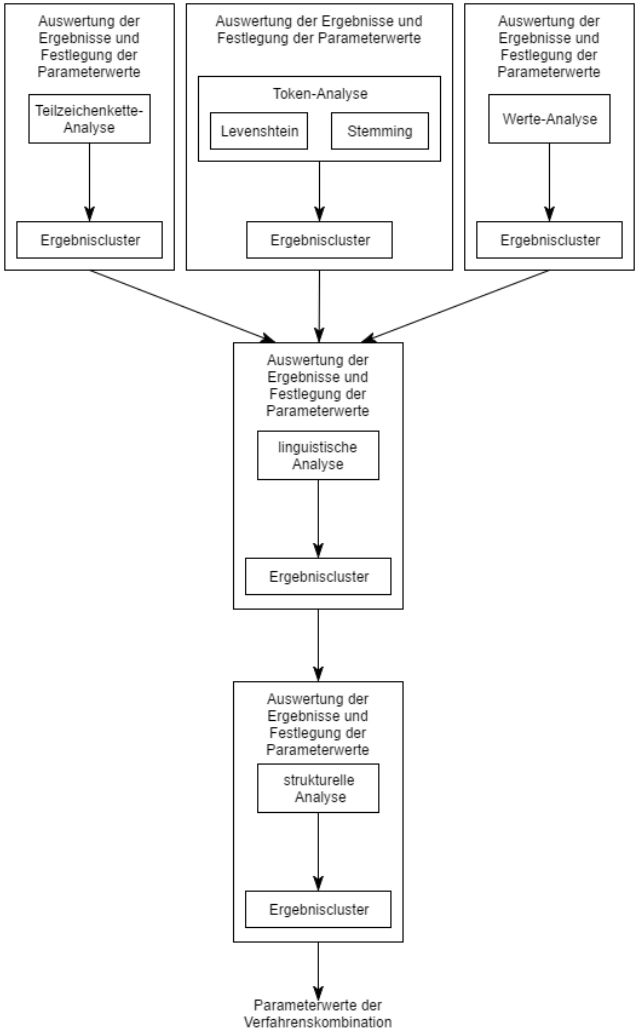
Hierarchisches Cluster-Verfahren

- Mithilfe der Ähnlichkeitswerte werden Cluster erstellt, welche semantisch gleiche Formularelemente enthalten
  - Hohe Ähnlichkeitswerte werden bevorzugt
  - Jedes Cluster enthält nicht mehr als ein Formularelement pro Dienstanbieter

Ähnlichkeiten		
Element 1	Element 4	0,95
Element 1	Element 5	0,9
Element 2	Element 5	0,7
Element 3	Element 4	0,8
Element 3	Element 6	0

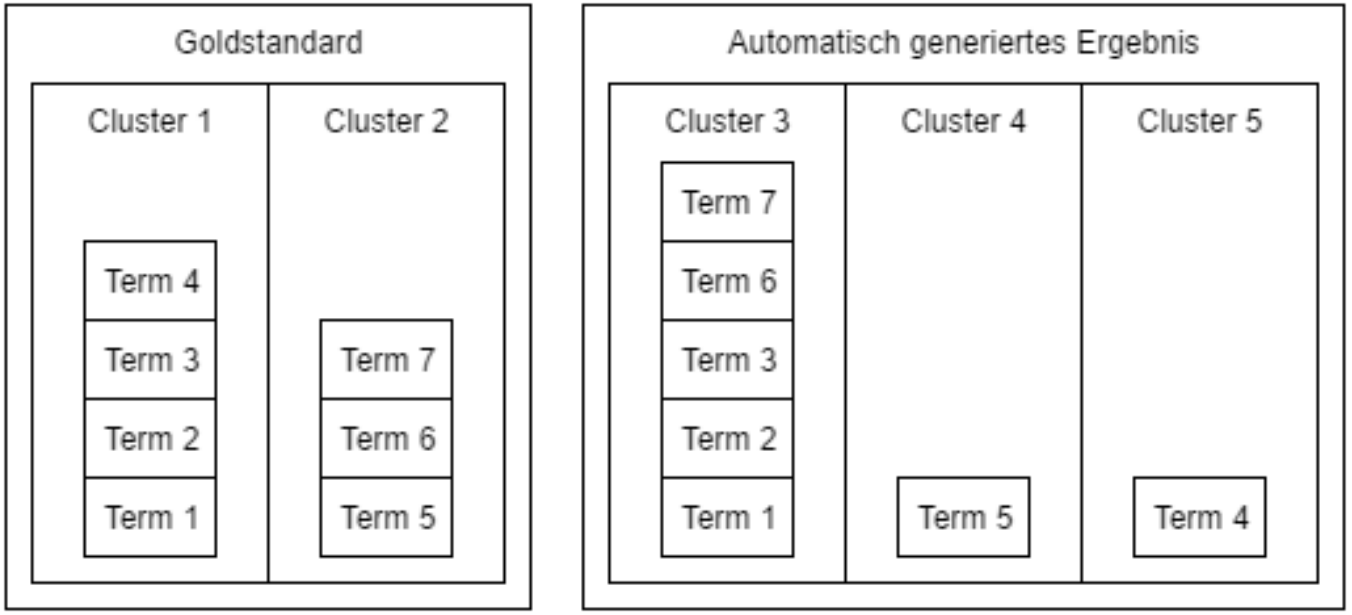



Objectgenerator
Analyzer
Calculator
Distributor
Clustergenerator
Matrixoperator
Konstructiongenerator
global
Objectgenerator


Auswertung der Ergebnisse und Festlegung der Parameterwerte
Teilzeichenkette-Analyse
Ergebniscluster
Auswertung der Ergebnisse und Festlegung der Parameterwerte
Token-Analyse
Levenshtein
Stemming
Ergebniscluster
Auswertung der Ergebnisse und Festlegung der Parameterwerte
Werte-Analyse
Ergebniscluster
Auswertung der Ergebnisse und Festlegung der Parameterwerte
linguistische Analyse
Ergebniscluster
Auswertung der Ergebnisse und Festlegung der Parameterwerte
strukturelle Analyse
Ergebniscluster
Parameterwerte der Verfahrenskombination

Verfahrenskombinationen

Verfahrens-kombination	Ln	Tk	Lv	St	Sb	Wa	Sa
1	x	x	x	x	x	x	x
2	x	x	x		x	x	x
3	x	x		x	x	x	x
4	x	x			x	x	x
5	x	x	x	x	x		x
6	x	x	x		x		x
7	x	x		x	x		x
8	x	x			x		x
9	x	x	x	x		x	x
10	x	x	x			x	x
11	x	x		x		x	x
12	x	x				x	x
13	x	x	x	x			x
14	x	x	x				x
15	x	x		x			x
16	x	x					x
17	x				x	x	x
18	x				x		x
19	x					x	x
20	x	x	x	x	x	x	
21	x	x		x	x	x	
22	x	x	x		x	x	
23	x	x			x	x	
24	x	x	x	x	x		
25	x	x	x		x		
26	x	x		x	x		
27	x	x			x		
28	x	x	x	x		x	
29	x	x	x			x	
30	x	x		x		x	
31	x	x				x	
32	x	x	x	x			
33	x	x	x				
34	x	x		x			
35	x	x					
36	x				x	x	
37	x				x		
38	x					x	


Goldstandard
Cluster 1
Cluster 2
Term 4
Term 3
Term 2
Term 1
Term 7
Term 6
Term 5
Automatisch generiertes Ergebnis
Cluster 3
Cluster 4
Cluster 5
Term 7
Term 6
Term 3
Term 2
Term 1
Term 5
Term 4

Erstellung globaler Objekte

- Bestimmung des Formularelementes
  - Wertebereichstypen
    - Endlich
      - Nur bestimmte Eingaben
    - Beschränkt
      - Eingaben innerhalb eines bestimmten Wertebereiches
    - Unendlich
      - Alle Eingaben sind zulässig

Endlich < Unendlich < Beschränkt

Hinweise

- http://sdqweb.ipd.kit.edu/wiki/Vortragshinweise

Grobe Struktur

- Ganz kurze Motivation
- keine Gliederung oder Ähnliches
- Verwandte Arbeiten (inklusive [Verweise] ins Backup)
- Ansatz
- Durchführung
- Evaluation
- Fazit

- Backup-Folien (hier kommt alles rein, was wir für den Vortrag nicht brauchen)
  - Besonders kniffelige Fragen an einem Beispiel erklären
  - Details, die zu tief gehen, aber interessant sein könnten
  - Folien, die beim Probevortrag rausgefallen sind

Literatur

- [Verweis2001] Eine schöne Referenz, IPD-Verlag, 2012.

Fazit und Ausblick

- Fazit des Werkzeuges
  - 70% der Abbildungen werden gefunden
  - 84% der gefundenen Abbildungen sind korrekt
- Fazit der Verfahren
  - Es existieren unterstützende Verfahren
  - Wichtige Verfahren:
    - Token-Analyse
    - Teilzeichenketten-Analyse
- Ausblick
  - Verwendung von Wörterbüchern
  - Verwendung von Synonym-Tabellen
  - Erstellung von komplexen Abbildungen

Linguistische Analyse

- Bestandteile
  - Token-Analyse
    - Führt eine Analyse mit beschreibenden und namensgebenden Attributen anhand ihrer Wörter durch
  - Teilzeichenketten-Analyse
    - Führt eine Analyse anhand der größten gemeinsamen Teilzeichenkette von beschreibenden und namensgebenden Attributen durch
  - Werte-Analyse
    - Führt eine Analyse anhand der möglichen Eingabewerte der Formularelemente durch

2do

- Fußzeile anpassen
  - Titel der Arbeit | Dein Name
  - Datum des Vortrags fest eintragen
- Einen schönen Vortrag bauen
  - Zielgruppe bedenken
  - Zeitvorgabe beachten
- 3x üben (vor dem Probevortrag)
- 3x üben (nach dem Probevortrag)
- Vortrag halten
- Feiern