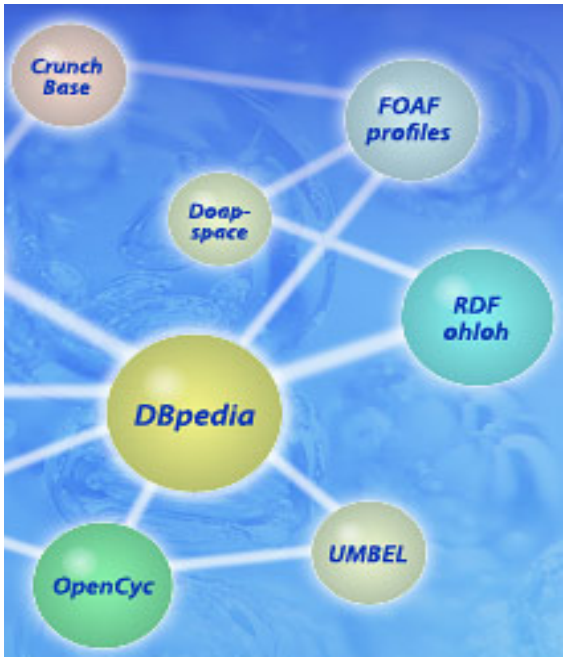
Profiling Linked Open Data with ProLOD

LiDDM: A Data Mining System for Linked Data



Seminar aus maschinellem Lernen

Frederik Janssen, Dr. Heiko Paulheim

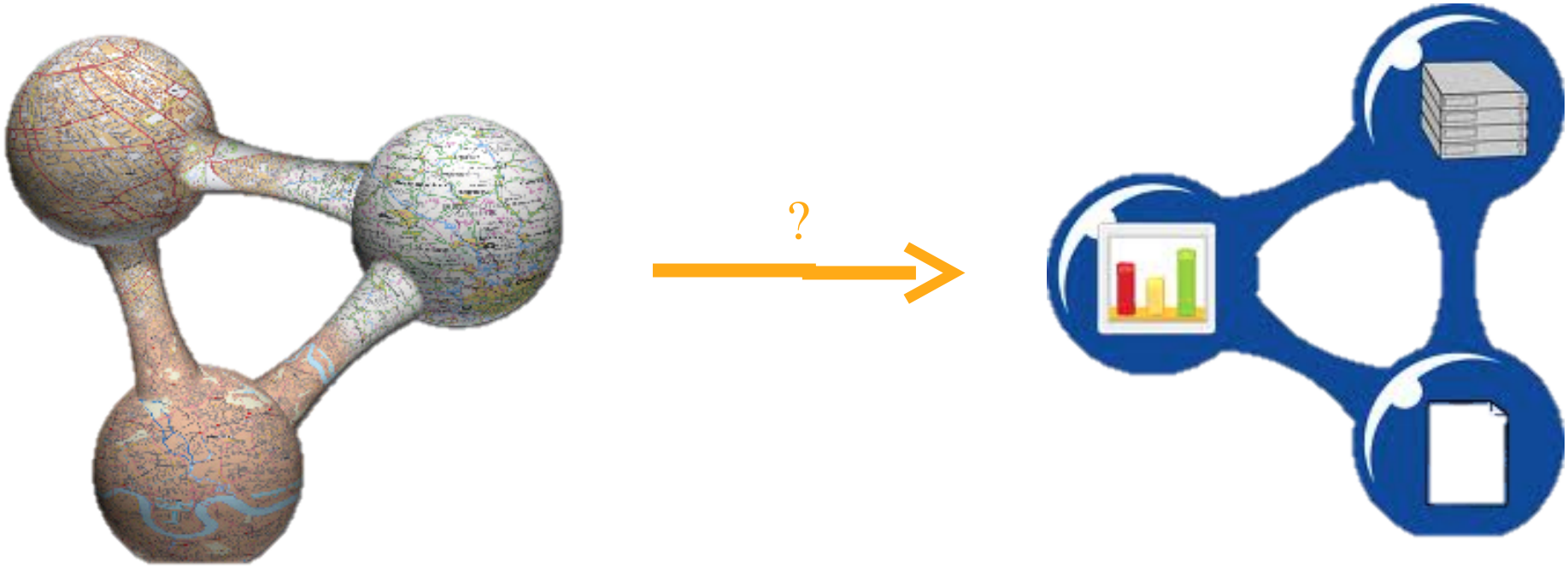
Profiling Linked Open Data with ProLOD

1. Linked Open Data
2. Profiling
3. Clustering and Labeling
4. Datatypes and Patterns

LiDDM: A Data Mining System for Linked Data

1. Einführung
2. Linked Data Data Mining Modell
3. Tool Umgebung
4. Ergebnisse

Linked Open Data



Traditionelle Profiling-methoden reichen nicht aus.

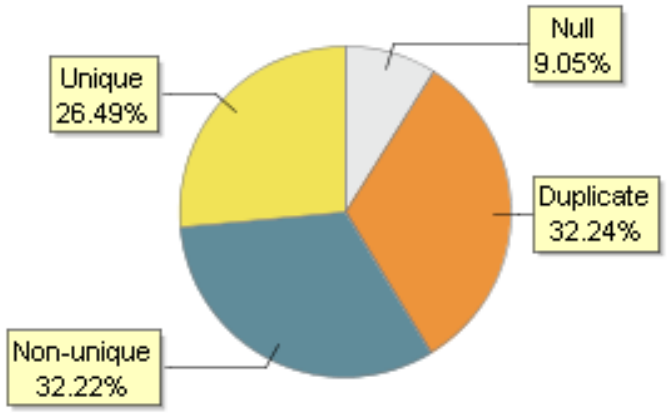
Lösung : Web-basiertes Program **: ProLOD**

Erkennung einer Semantik von Spalten

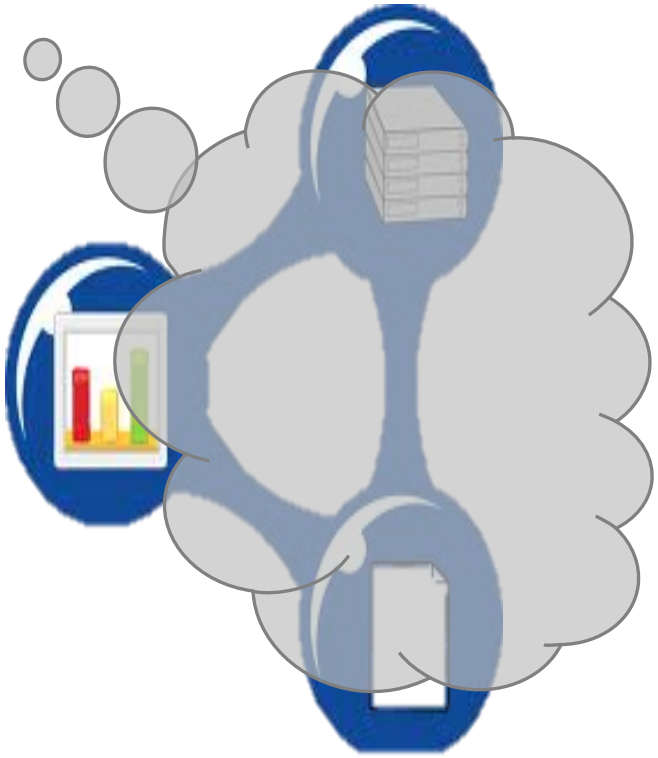
Expression	Type	Domain	Non-null	Null	Unique	Distinct
src_name	string	specval pa...	139,317	0	130,464	132,575
src_gender	string	enum	15,780	123,537	0	4
src_birth_d...	string	day pattern	124,340	14,977	2,798	25,018
src_sin	string	pattern	134,830	4,487	134,766	134,798
src_card	string	long	55,904	83,413	15,800	35,630
src_address	string	enum patt...	9	139,308	5	7
src_primar...	integer		139,317	0	139,315	139,316

Zweck:

- Erkennung von einzigartigen Variablen
- Erstellung von Patterns



Profiling-LOD Methode



gruppieren

Analyse einfacher Statistiken

Verfeinerung des Profilings

Proof-of-Concept

Ohne Ontologie

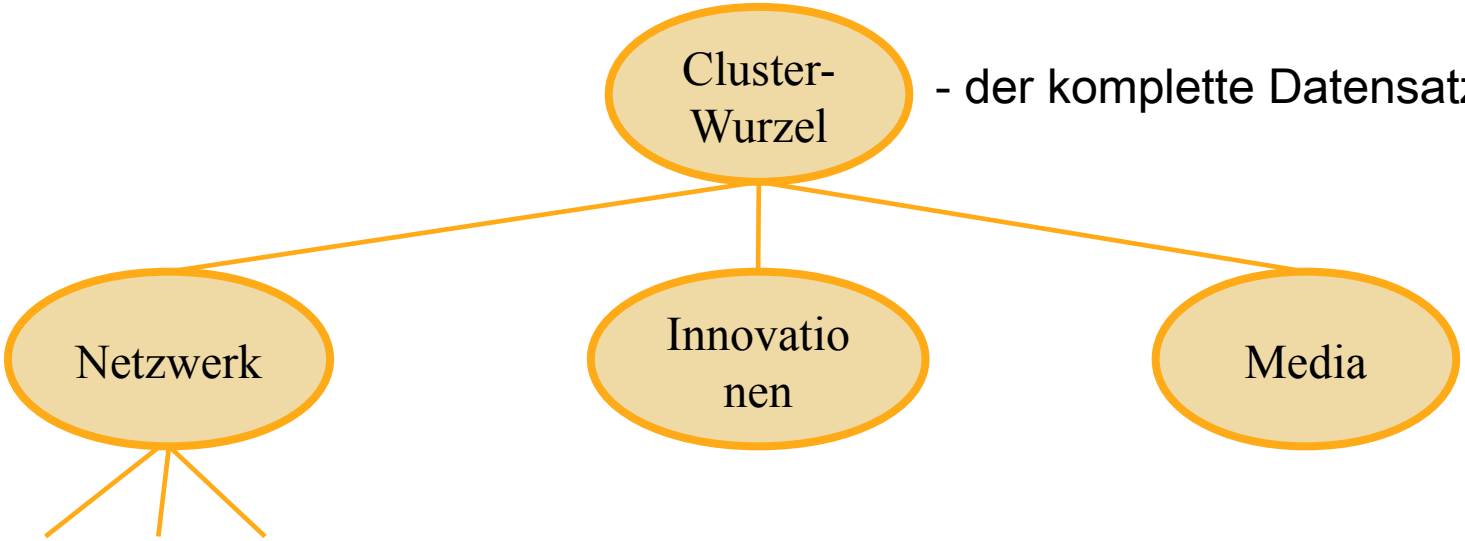
und

34,2 Mil. Tripeln

Clustering

Ein gutes Clustering: semantikabhängig

Probleme: - die Clusteranzahl ist nicht bekannt.
- die Größe der Datasätze



Strategie bei einer großen Anzahl der Tripeln :

Wurzelcluster
im Vorbereitungsschritt erstellen

Die Grenzwerte der Unähnlichkeiten
gelten zur Bildung eines neuen Clusters

K-Means Algorithmus

Labeling

TF-IDF um die m-wichtigsten Terme zu berechnen.

Beispiel für m = 3

Label	Sample subjects
minister politician mayor	Angela_Merkel Ted_Kennedy Presidency_of_George_Washington
film directed starring	Titanic_(1997_film) Metropolis_(film) Frankenstein_(1910_film)
club football league	FC_Bayern_Munich Liverpool_F.C. Los_Angeles_Galaxy

Für einen besseren Zusammenhang und
für das Verständnis der Tripelsätze findet man

- die assoziativen Attribute
- die Assoziationsregeln im Cluster.

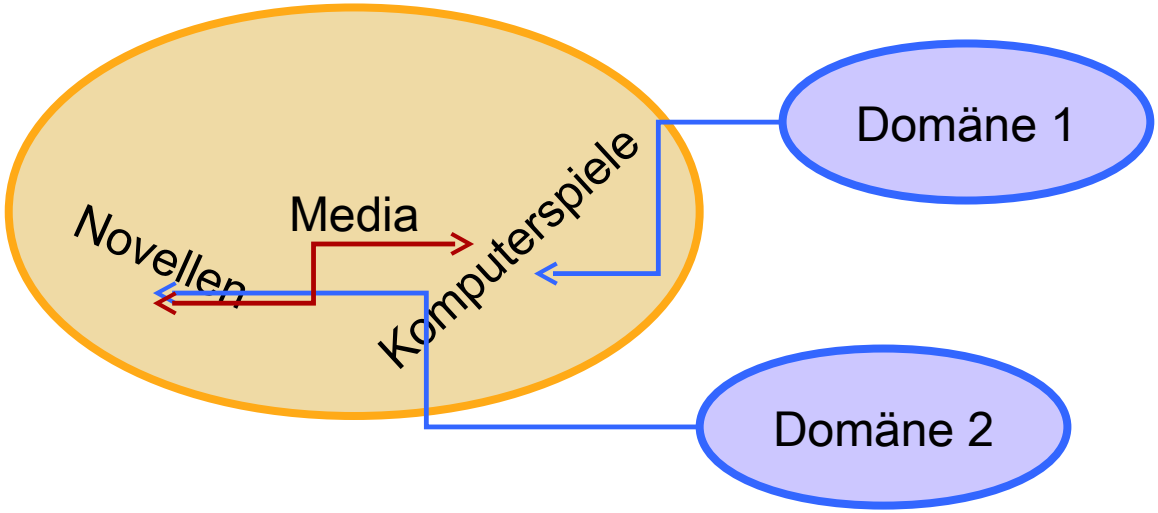
Als Ergebniss für ein Buch in “Media” kommen häufig author, isbn, genre vor.

Rule	Confidence	Correlation Coefficient
$genre, isbn \Rightarrow author$	0.99	0.67
$isbn \Rightarrow author$	0.92	0.66
$isbn \Rightarrow author, genre$	0.83	0.66
$author, genre \Rightarrow isbn$	0.70	0.66
$author \Rightarrow isbn$	0.64	0.66
$author \Rightarrow genre, isbn$	0.58	0.67

Das Modell ist in der Lage, sowohl die possitiven, als auch die negativen Assoziationsregeln zu erkennen, um die Abhängigkeiten zwischen Prädikaten in einem Cluster zu finden.

Zweite Möglichkeit : negative Abhängigkeiten.

Equivalente Semantik



author =>¬ developer und developer =>¬ author : Konfidenz 90%

Genauso mit name und title

X hat eine Verbindung zur Y durch Prädikat A

Je mehr Einheiten durch A und B verbunden sind, desto höher ist der Zusammenhang der inversen Prädikatenpaare.

z.B. für das Buch “Into the Wild” :

Into the Wild —debutWorks→ Jon Krakauer

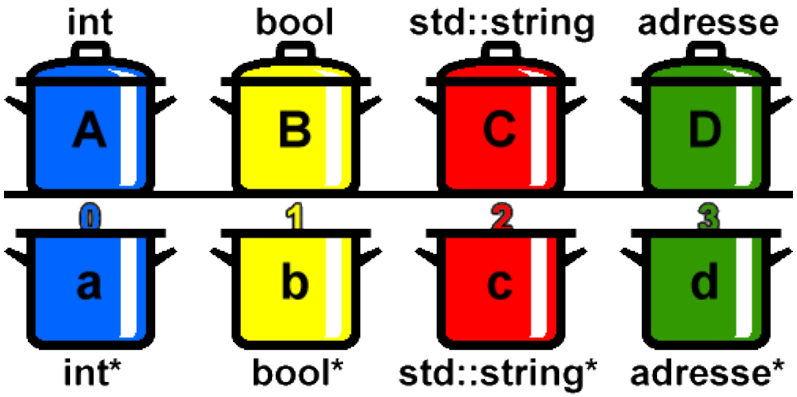
Jon Krakauer —author→ Into the Wild

Tabelle der inversen Linkpaare :

$\overrightarrow{PredicateA}$	$\overleftarrow{PredicateB}$	Corr Coef	Frequency
before	after	0.239	28856
sisterStations	sisterStations	0.749	7494
precededBy	followedBy	0.830	7097
spouse	spouse	0.322	1964
before	before	-0.003	738
star	exoplanet	0.895	188

Wichtig für das semantische (Pre)-Clustering.

Variablenverteilung in den Datentypen



Im Bereich der Datentypen verteilt

Patterns

Normalisierte Patterns

ProLOD - Tool

Preprocessing

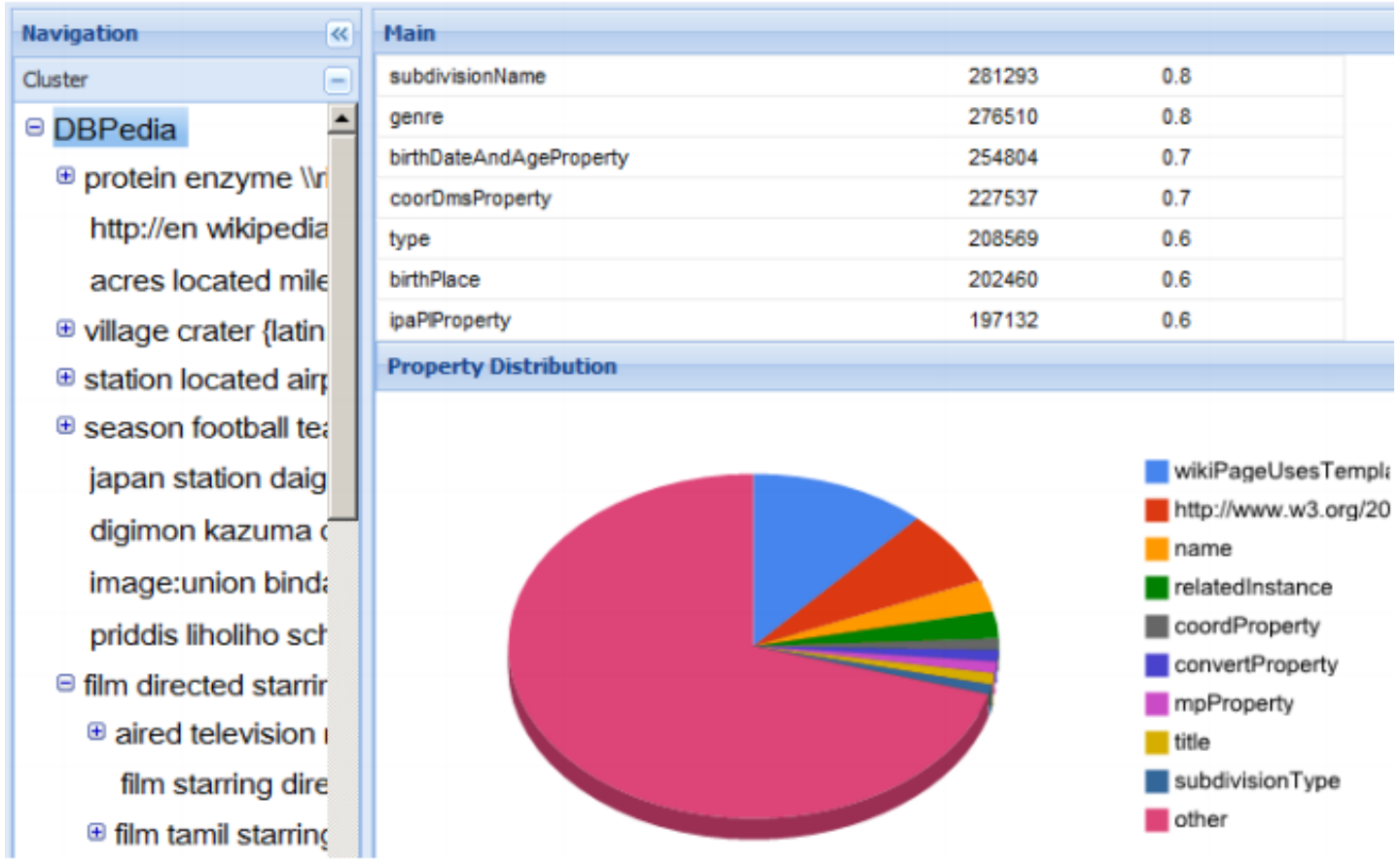
1. Tripel umgewandelt in der Datenbank gespeichert

2. Clustering & Labeling

3. Jedem Tripel ein Pattern bzw. Datentyp zugewiesen

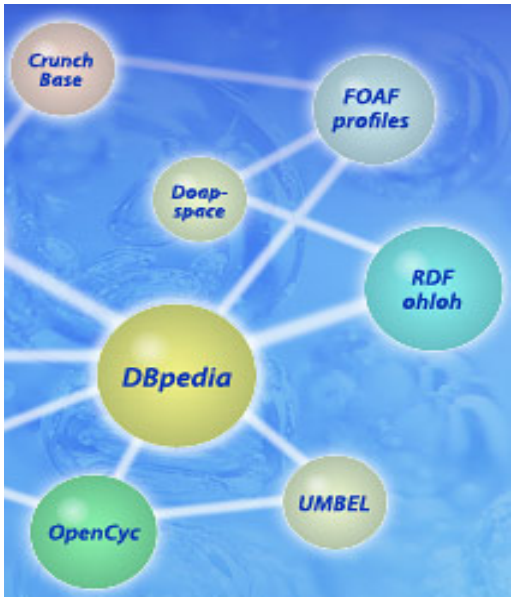
Aufwand : 1 Tag mit DBPedia

Realtime Profiling



A Data Mining System for Linked Data

extrem wachsende verlinkte Daten:

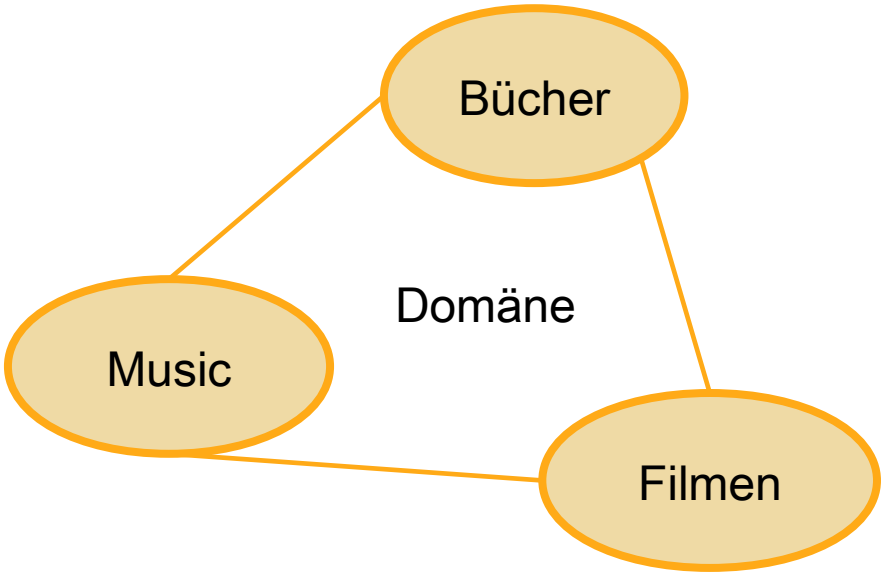


- Sammeln von verschiedenen Quellen
- Integieren
- Für Statistiken verfeinern

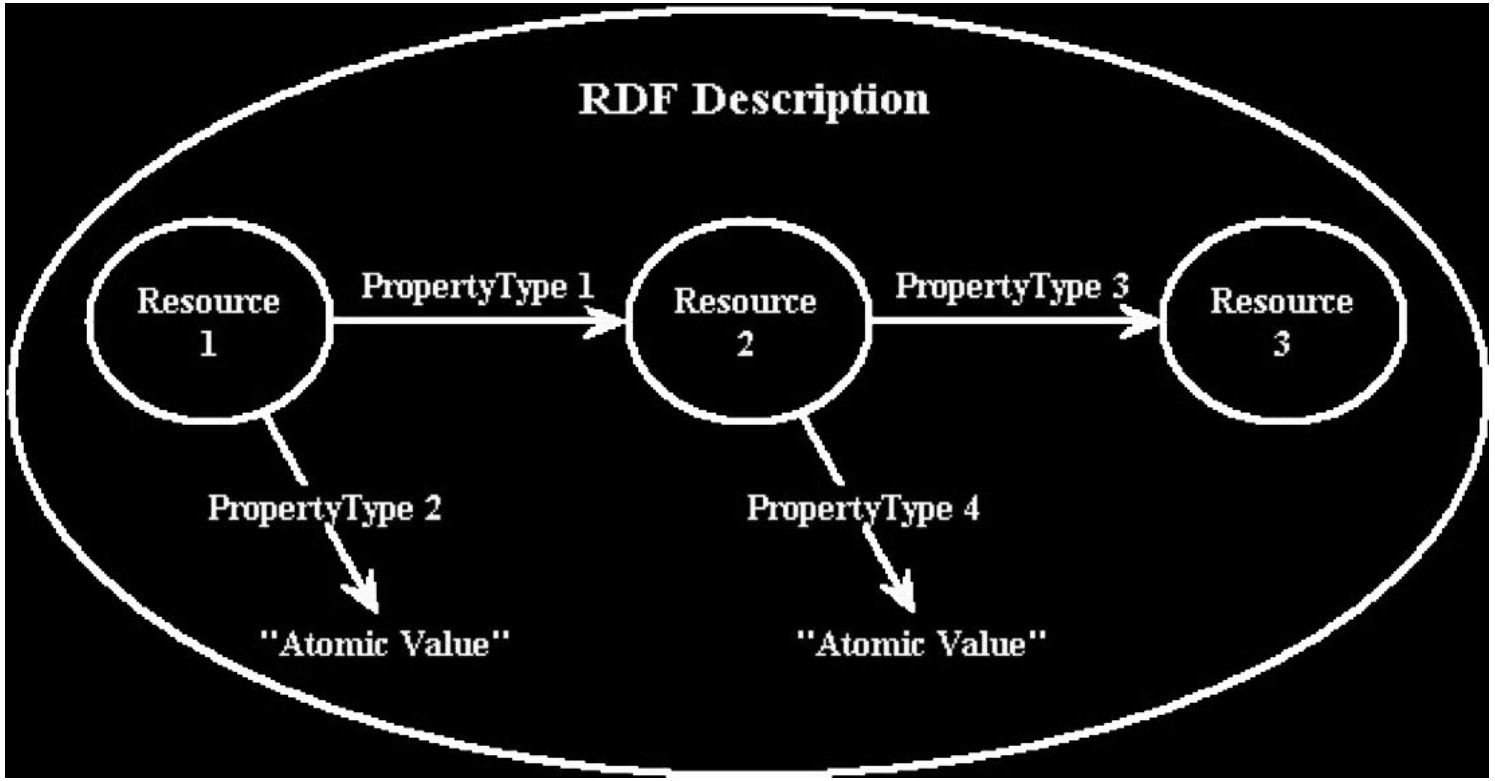
Deswegen hat man dafür ein Modell entwickelt

komplex und dynamisch.

- Fundament für Linked Data bei Web
- verknüpft die Daten aus verschiedenen Domänen



- bietet verschiedene Datasätze in RDF-Format:



Wordl FactBook

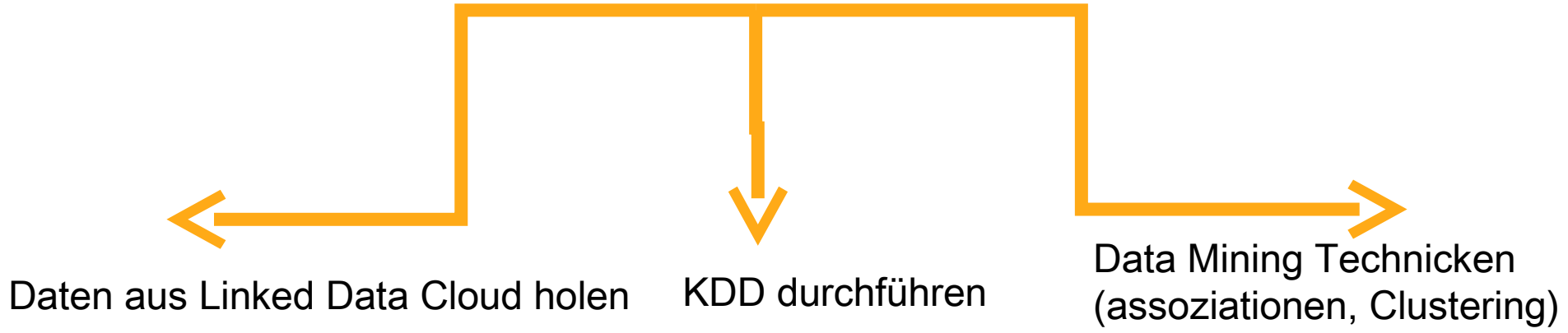
Data.gov

DBPedia

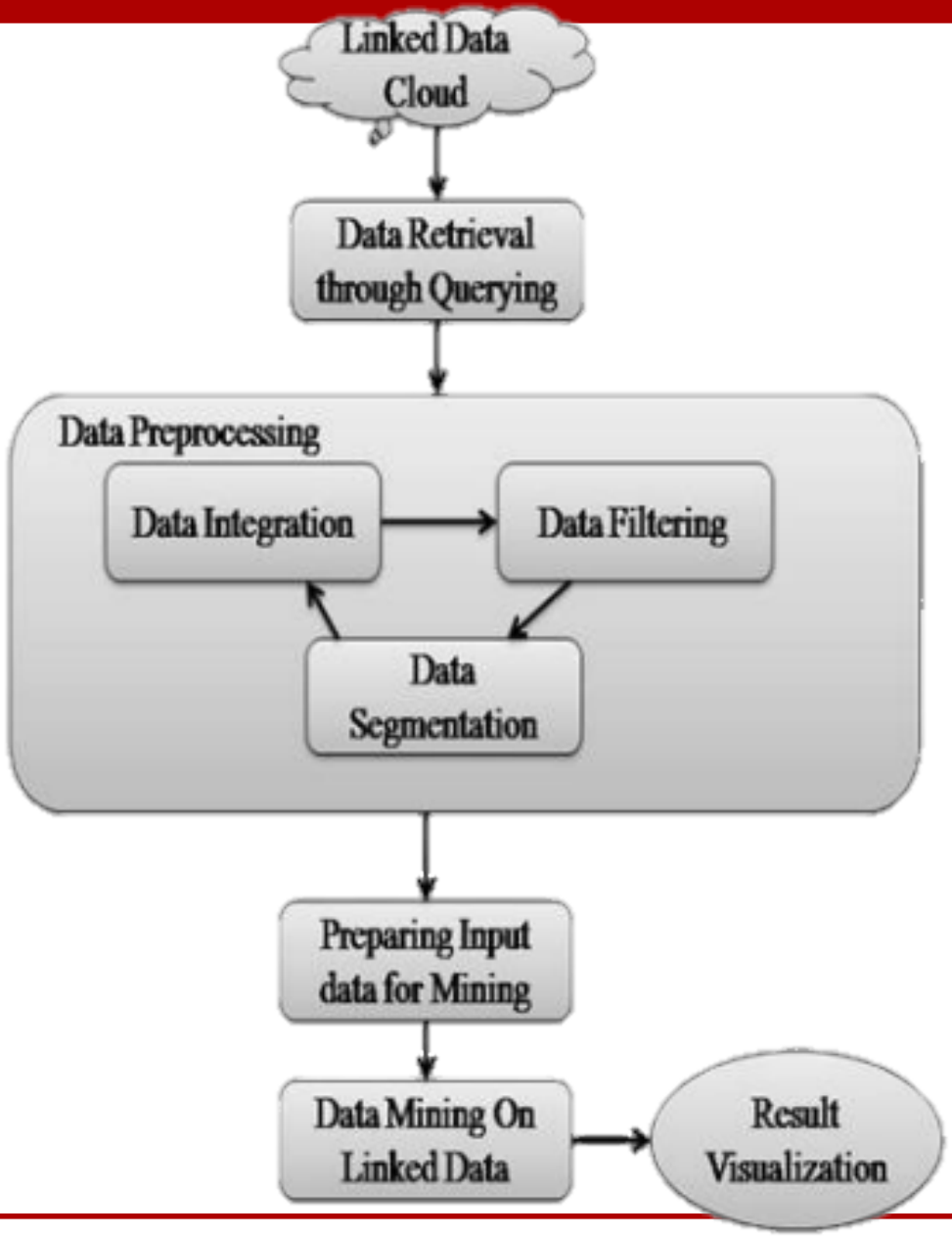
- Informationen aus der reellen Welt

- Vorhersage für künftige Statistiken

Die wertvollen, versteckten Informationen extrahieren.



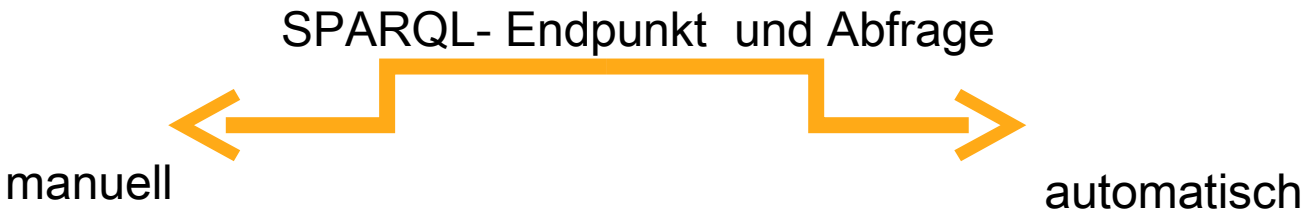
visualisieren


Linked Data Cloud
Data Retrieval through Querying
Data Preprocessing
Data Integration
Data Filtering
Data Segmentation
Preparing Input data for Mining
Data Mining On Linked Data
Result Visualization

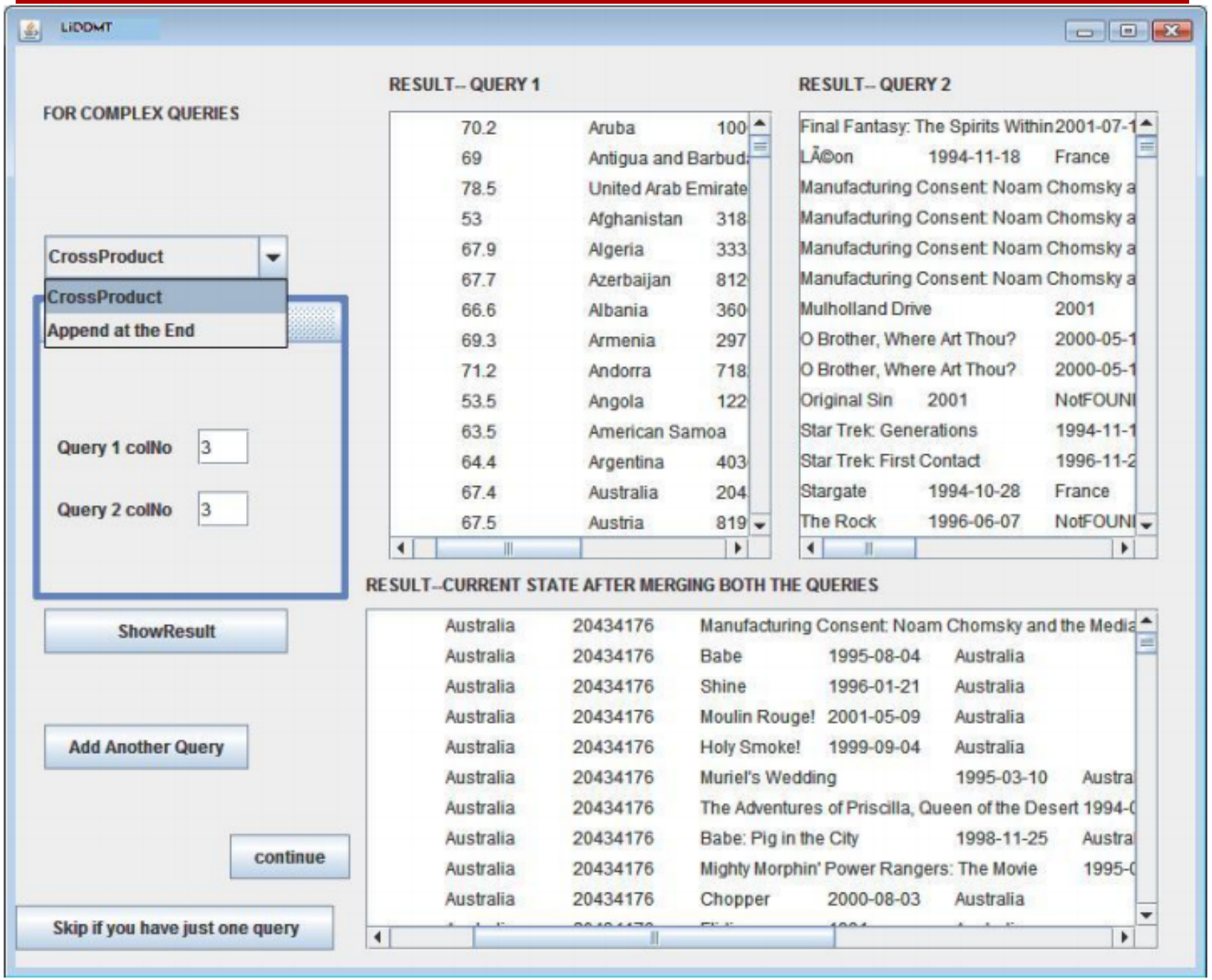
Tool Umgebung

Linked Data Data Mining Tool - LiDDMT

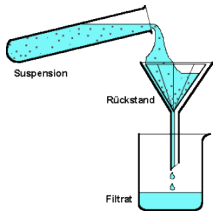
Schritt 1.



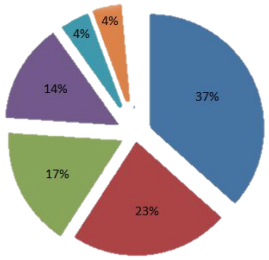
Abfrage Builder liefert die möglichen Prädikaten und eine spezifizierte Abfrage.


LiDDMT
FOR COMPLEX QUERIES
RESULT-- QUERY 1
RESULT-- QUERY 2
CrossProduct
CrossProduct
Append at the End
Query 1 colNo
3
Query 2 colNo
3
ShowResult
Add Another Query
continue
Skip if you have just one query
RESULT--CURRENT STATE AFTER MERGING BOTH THE QUERIES

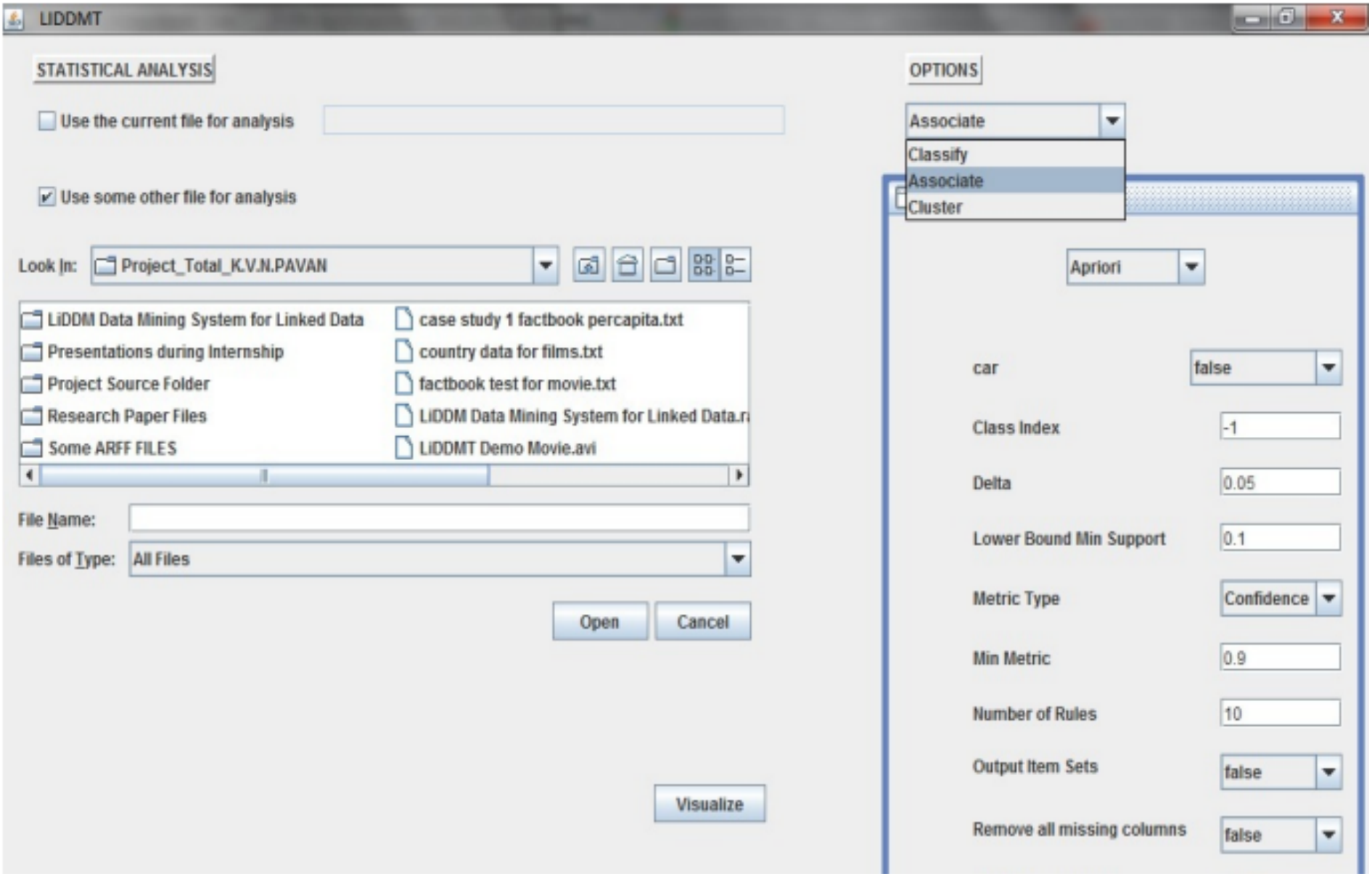
Schritt 3. Fillterung



Schritt 4.



Schritt 5. Die Daten im geeigneten Format für Mining speichern :
ARFF – Attribute-Relation File Format


LIDDMT
STATISTICAL ANALYSIS
Use the current file for analysis
Use some other file for analysis
Look In:
Project_Total_K.V.N.PAVAN
LiDDM Data Mining System for Linked Data
Presentations during Internship
Project Source Folder
Research Paper Files
Some ARFF FILES
case study 1 factbook percapita.txt
country data for films.txt
factbook test for movie.txt
LiDDM Data Mining System for Linked Data.r
LiDDMT Demo Movie.avi
File Name:
Files of Type:
All Files
Open
Cancel
Visualize
OPTIONS
Associate
Classify
Associate
Cluster
Apriori
car
false
Class Index
-1
Delta
0.05
Lower Bound Min Support
0.1
Metric Type
Confidence
Min Metric
0.9
Number of Rules
10
Output Item Sets
false
Remove all missing columns
false

Ergebnisse

World FactBook für Bildungsniveau der Arbeitskräfte in der Landwirtschaft, in der Wirtschaft und im Services-Bereich.

Attribute sind in drei Klassen segmentiert;

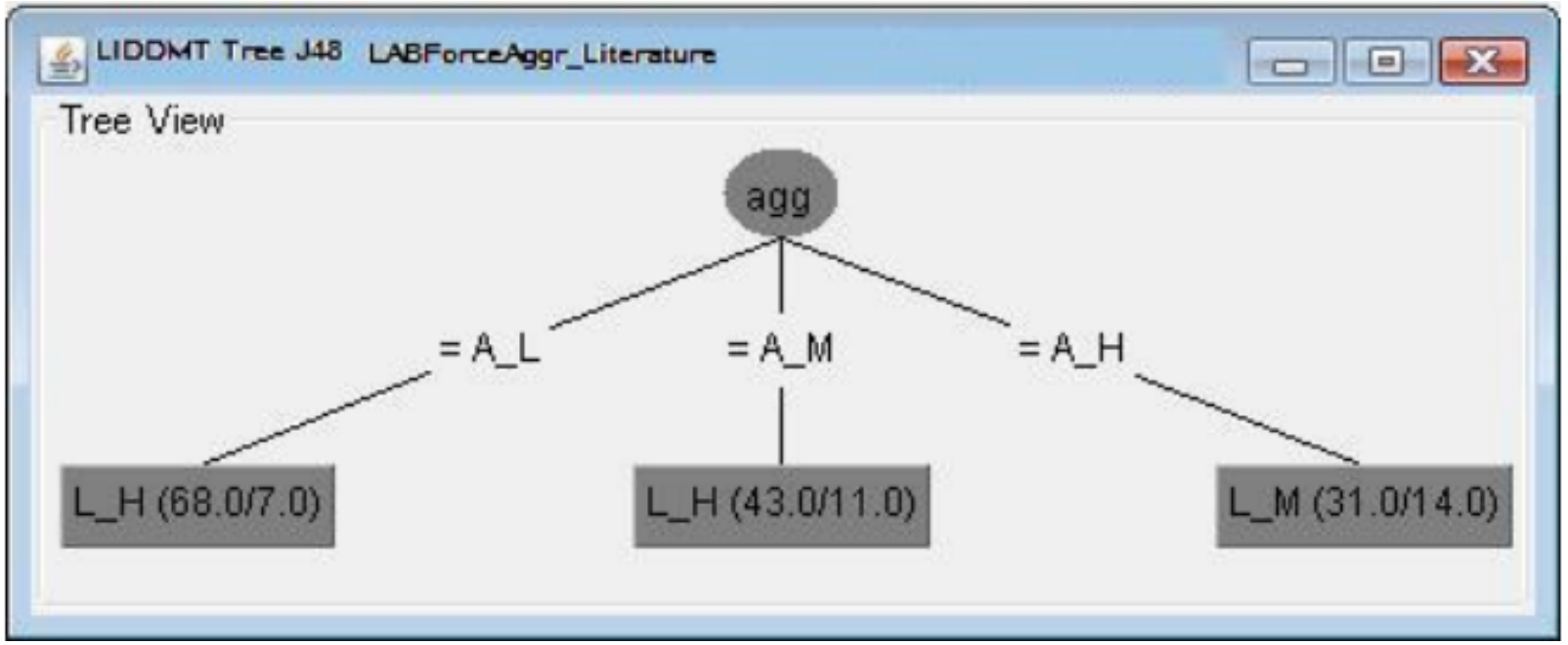
A_L: agriculture low
A_M: agriculture medium
A_H: agriculture high.

Bildungsniveau in drei Klassen segmentiert:

low 0 - 50

medium 50 - 85

high 85 - 100


LIDDMT Tree J48 LABForceAggr_Literature
Tree View
agg
= A_L
= A_M
= A_H
L_H (68.0/7.0)
L_H (43.0/11.0)
L_M (31.0/14.0)

Danke für Ihre Aufmerksamkeit!

TECHNISCHE UNIVERSITÄT DARMSTADT

Autoren:

“Profiling Linked Open Data with ProLOD” : Christoph Böhm, Felix Naumann, Ziawasch Abedjan, Dandy Fenz, Toni Grütze, Daniel Hefenbrock, Matthias Pohl, David Sonnabend

“LiDDM: A Data Mining System for Linked Data” : Venkata Narasimha Pavan Kappara, Ryutaro Ichise , O.P. Vyas

Referenzen:

http://www.highscore.de/cpp/einfuehrung/zeiger.html
http://ex-ample.blogspot.com/2011/07/tthe-weka-environment-weka-is.html
http://jena.sourceforge.net/