

# Oberseminar Data Mining

Systeme und Tools zum Data Mining: RapidMiner



# Motivation



## Inhalt

- 1 Das Projekt RapidMiner
- 2 Funktionen
- 3 KDD-Prozess
- 4 Weitere Werkzeuge von Rapid-I
- 5 Zusammenfassung

## Entwicklung

- entwickelt an der Technischen Universität Dortmund
- erschienen im Jahre 2001
- anfangs unter dem Namen YALE - ("Yet Another Learning Environment") veröffentlicht
- 2007 umbenannt in RapidMiner
- zu diesem Zeitpunkt Version 4.0
- seit Februar 2010 Version 5.0

## Produktübersicht I

- lizenziert unter der AGPL bzw. proprietär
- Open-Source-Software
- erhältlich in der Community- oder Enterprise Edition
- komplett in Java geschrieben und damit auf allen großen Plattformen lauffähig
- bietet die Möglichkeit über Java API von externen Programmen genutzt zu werden

## Produktübersicht II

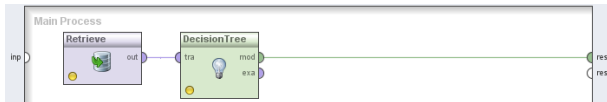
- Umgebung zum maschinellen Lernen und zur Umsetzung des KDD-Prozesses (insbesondere des Data Minings)
- Realisierung mittels einer Reihe von Operatoren (z.Z. ca. 500 verschiedene), z.B.:
  - Algorithmen zum maschinellen Lernen
  - Datenvorverarbeitungsoperatoren
  - Meta-Operatoren
  - Operatoren zur Visualisierung
  - Operatoren zum Im- und Export
  - ...
- RapidMiner nutzt XML um Operatorbäume darzustellen, die den KDD-Prozess modellieren

## XML-Operatorbaum

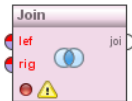
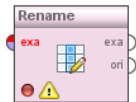
```

1 <?xml version="1.0" encoding="UTF-8" standalone="no" ?>
2 <process version="5.0">
3   <!-- [...] -->
4   <operator activated="true" class="process" expanded="true" name="Root">
5     <process expanded="true" height="541" width="675">
6       <operator activated="true" class="retrieve" expanded="true" height="60
7         " name="Retrieve" width="90" x="45" y="30">
8         <parameter key="repository_entry" value="../../../ data/ Golf" />
9       </operator>
10      <operator activated="true" class="decision_tree" expanded="true"
11        height="76" name="DecisionTree" width="90" x="180" y="30" />
12      <connect from_op="Retrieve" from_port="output" to_op="DecisionTree"
13        to_port="training_set" />
14      <connect from_op="DecisionTree" from_port="model" to_port="result_1" />
15    <!-- [...] -->
16  </process>
17 </operator>
18 </process>

```

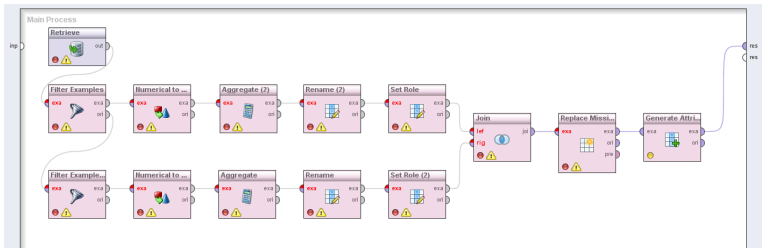


## Beispiel-Operatoren





## Beispiel-Operatorkette



# Überblick

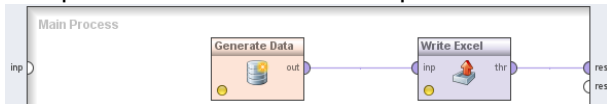
- 1 Das Projekt RapidMiner
- 2 Funktionen**
- 3 KDD-Prozess
- 4 Weitere Werkzeuge von Rapid-I
- 5 Zusammenfassung

## Schnittstellen

- 3 Möglichkeiten zur Bedienung:
  - Server Modus (Kommandozeile)
  - über Java API aus externen Programmen
  - GUI Modus

## Server Modus

- Voraussetzungen:
  - Umgebungsvariable „RAPIDMINER\_HOME“ auf Verzeichnis der Installation setzen
  - optional: PATH-Variable zu „rapidminer.bat“ setzen
- Beispiel: Datei TestProcess.rmp



- Aufruf allg.: `rapidminer [-f] Prozessname`
- am Beispiel: `rapidminer -f TestProcess.rmp`

## Einbindung in externes Programm

```
import com.rapidminer.Process;
import com.rapidminer.RapidMiner;
import com.rapidminer.operator.Operator;
import com.rapidminer.operator.OperatorException;
import com.rapidminer.operator.generator.ExampleSetGenerator;
import com.rapidminer.tools.OperatorService;

public class ProcessCreator {

    public static void main(String[] argv) {
        Process process = createProcess();
        System.out.println(process.getRootOperator().createProcessTree(0));

        try {
            process.run();
        } catch (OperatorException e) {
            e.printStackTrace();
        }
    }

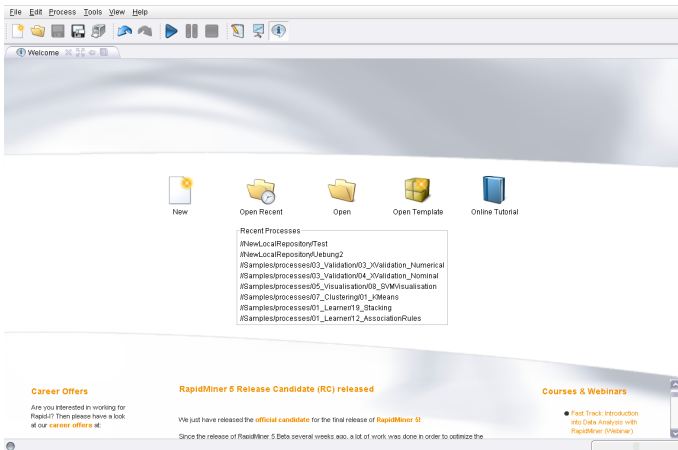
    // [...]
```

## Einbindung in externes Programm

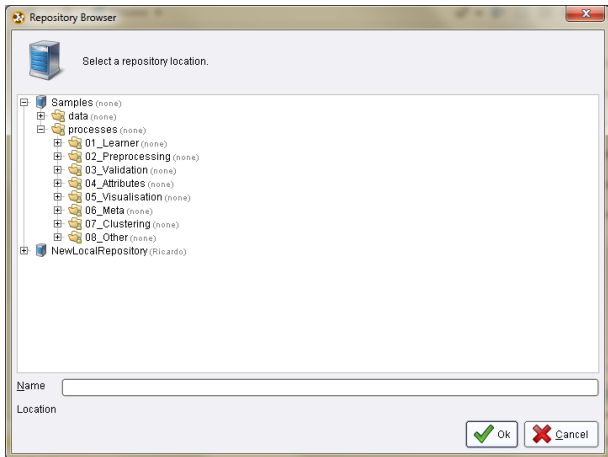
```
// [...]  
  
public static Process createProcess() {  
    RapidMiner.init();  
    Process process = new Process();  
    try {  
        Operator inputOperator = OperatorService.createOperator(  
            ExampleSetGenerator.class);  
        inputOperator.setParameter("target_function", "sum_classification");  
        process.getRootOperator().getSubprocess(0).addOperator(inputOperator);  
    } catch (Exception e) { e.printStackTrace(); }  
    return process;  
}  
  
/* Ausgabe:  
 *  
 * Process[0] (Process)  
 * subprocess 'Main Process'  
 * +- Generate Data[0] (Generate Data)  
 */  
}
```

(Quelle: nach [RI09])

## Oberfläche - Start

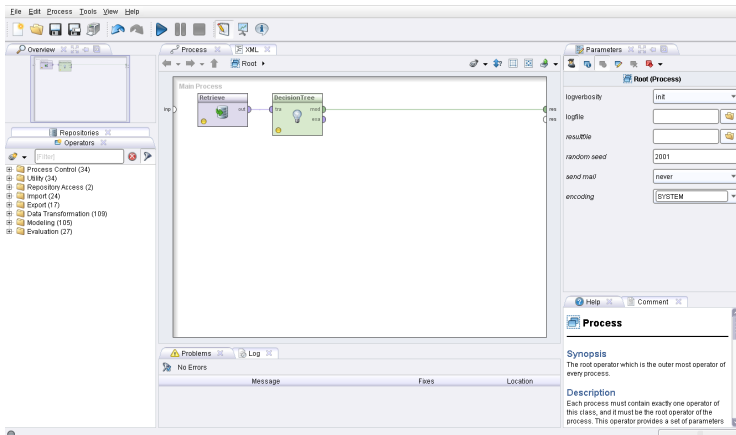


## Oberfläche - Neuer Prozess





## Oberfläche - Design Workspace



The screenshot displays the RapidMiner Design Workspace. The main area shows a process flow starting with a 'Retrieve' operator (purple box) connected to a 'Decision Tree' operator (green box). The 'Retrieve' operator has an 'out' port connected to the 'Decision Tree' operator's 'in' port. The 'Decision Tree' operator has an 'out' port connected to the 'Root (Process)' operator's 'in' port. The 'Root (Process)' operator is a large white box with an 'out' port on the right side.

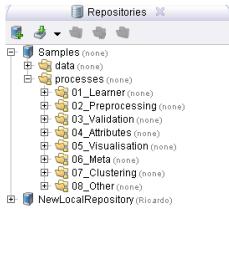
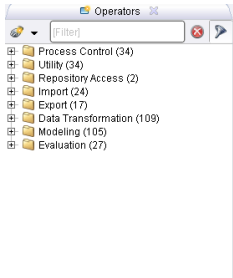
On the left side, there is a 'Repositories' panel with a search bar and a list of categories: Process Control (34), UNITY (34), Repository Access (2), Import (24), Export (17), Data Transformation (109), Modeling (105), and Evaluation (27).

On the right side, there is a 'Parameters' panel for the 'Root (Process)' operator. The parameters are:
 

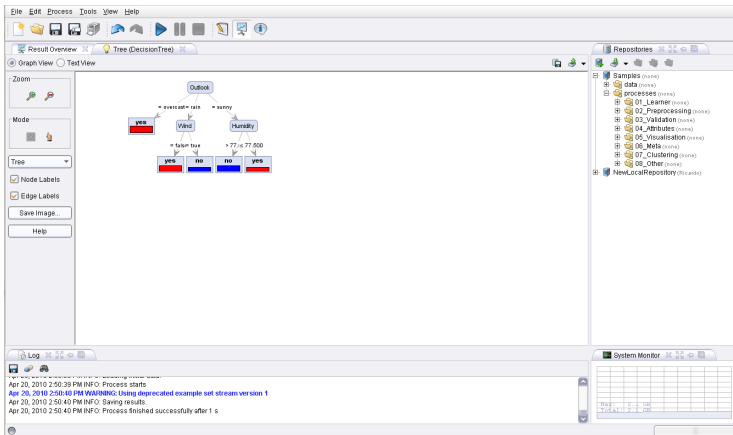
- logverbosity: int (dropdown menu)
- logfile: (text input field)
- resultfile: (text input field)
- random seed: 2001 (text input field)
- send mail: never (dropdown menu)
- encoding: SYSTEM (dropdown menu)

At the bottom, there is a 'Problems' panel showing 'No Errors' and a 'Log' panel. The status bar at the very bottom shows 'Message', 'Files', and 'Location'.

## Oberfläche - Operatoren und Repositories



# Oberfläche - Result Workspace



The screenshot displays the RapidMiner Result Workspace interface. The main area shows a decision tree visualization for a 'DecisionTree' process. The tree structure is as follows:

```


    graph TD
      Outlook -- "overcast = rain" --> Yes1[yes]
      Outlook -- "sunny" --> Wind[Wind]
      Outlook -- "sunny" --> Humidity[Humidity]
      Wind -- "false" --> Yes2[yes]
      Wind -- "true" --> No1[no]
      Humidity -- "false" --> No2[no]
      Humidity -- "true" --> Yes3[yes]
  
```



The interface includes a menu bar (File, Edit, Process, Tools, View, Help), a toolbar with icons for file operations and execution, and a 'Result Overview' tab. On the left, there are controls for 'Zoom', 'Mode', and 'Tree' view, along with checkboxes for 'Node Labels' and 'Edge Labels', and buttons for 'Save Image...' and 'Help'. On the right, a 'Repositories' panel shows a hierarchical view of data sources and processes, including 'Samples (data)', 'data (data)', 'processes (process)', and various numbered processes like '01\_Learner (process)', '02\_Preprocessing (process)', etc. At the bottom, a 'Log' window shows system messages, and a 'System Monitor' window displays resource usage statistics.

## Visualisierung

- es bestehen 3 Möglichkeiten der Visualisierung von Ergebnissen
  - Meta-Daten-Sicht (*Meta Data View*)
  - Daten-Sicht (*Data View*)
  - grafische Darstellungs-Sicht (*Plot View*)
- bei der grafischen Darstellung besteht die Möglichkeit diverse Visualisierung mittels 2D- und 3D-Grafiken zu erzeugen

## Beispiel - Meta Data View

Meta Data View
  Data View
  Plot View
 

ExampleSet (14 examples, 1 special attribute, 4 regular attributes)  

Role	Name	Type	Statistics	Range	Missings
label	Play	nominal	mode = yes (9), least = no (5)	no (5), yes (9)	0
regular	Outlook	nominal	mode = sunny (5), least = overcast (4)	rain (5), overcast (4), sunny (5)	0
regular	Temperature	integer	avg = 73.571 +/- 6.572	[64.000 ; 85.000]	0
regular	Humidity	integer	avg = 80.286 +/- 9.840	[65.000 ; 96.000]	0
regular	Wind	nominal	mode = false (8), least = true (6)	true (6), false (8)	0

## Beispiel - Data View

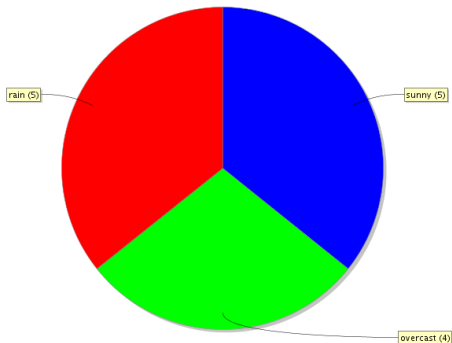
Meta Data View 
  Data View 
  Plot View

ExampleSet (14 examples, 1 special attribute, 4 regular attributes) 
 View Filter (14 / 14): all

Row No.	Play	Outlook	Temperature	Humidity	Wind
1	no	sunny	85	85	false
2	no	sunny	80	90	true
3	yes	overcast	83	78	false
4	yes	rain	70	96	false
5	yes	rain	68	80	false
6	no	rain	65	70	true
7	yes	overcast	64	65	true
8	no	sunny	72	95	false
9	yes	sunny	69	70	false
10	yes	rain	75	80	false
11	yes	sunny	75	70	true
12	yes	overcast	72	90	true
13	yes	overcast	81	75	false
14	no	rain	71	80	true

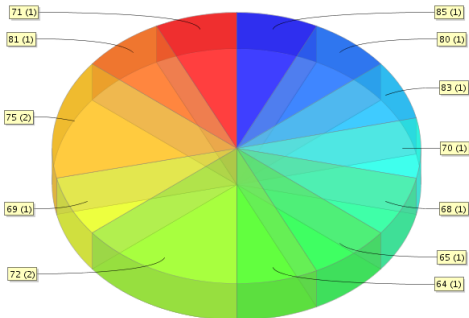
## Beispiel - Plot View (Pie)

● sunny (5) ● overcast (4) ● rain (5)



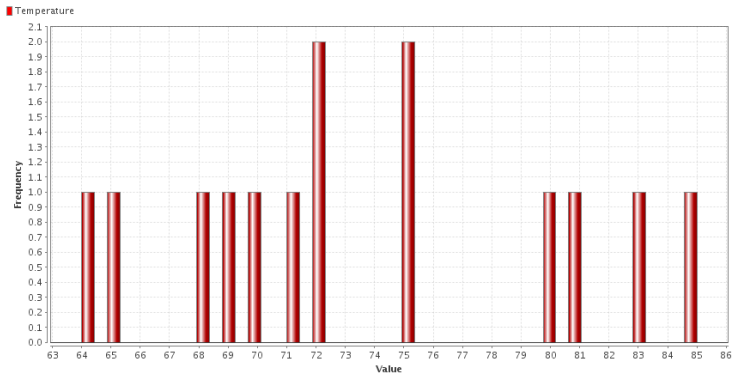
## Beispiel - Plot View (Pie 3D)

● 85 (1) ● 80 (1) ● 83 (1) ● 70 (1) ● 68 (1) ● 65 (1) ● 64 (1) ● 72 (2) ● 69 (1) ● 75 (2) ● 81 (1) ● 71 (1)





## Beispiel - Plot View (Histogramm)



## Erweiterungsmöglichkeiten

- RapidMiner bietet die Möglichkeit über Erweiterungen den Funktionsumfang zu vergrößern
- Beispiele für Erweiterungen sind:
  - Weka Extension
  - Parallel Processing
  - Text Processing
  - Web Mining
  - Reporting Extension
  - Series Processing
  - PMML

## Datenformate

Funktion	Formate
Import	CSV, Excel, Access, BibTeX, Database, DBase, URL, SPSS, AML, ARFF, XRFF, Stata, Sparse, C4.5, DasyLab
Export	CSV, Excel, Access, AML, ARFF, XRFF, Database

# Überblick

- 1 Das Projekt RapidMiner
- 2 Funktionen
- 3 KDD-Prozess**
- 4 Weitere Werkzeuge von Rapid-I
- 5 Zusammenfassung

## Wiederholung KDD-Prozess

- KDD = Knowledge Discovery in Databases
- Schritte:
  - ① Datenselektion und -extraktion
  - ② Datenbereinigung und -transformation
  - ③ Data Mining
  - ④ Interpretation
- Umsetzung in RapidMiner als Operatoren-Kette

## Funktionsweise im RapidMiner

- Austausch von IOObjects zwischen Operatoren
- Datenmenge als ExampleSet bezeichnet
  - entspricht Tabelle
  - Examples sind die Zeilen
  - Attribute sind die Spalten

# Attribute

## Rollen

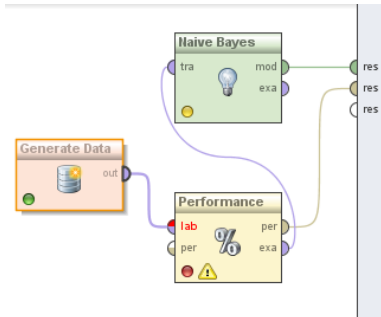
- regular attributes
- special attributes
  - ID
  - Label
  - Prediction
  - Cluster
  - Weight
  - Batch

## Typen

- (bi-/poly-)nominal
- numeric
- date
- text

## Farbliche Markierung in GUI

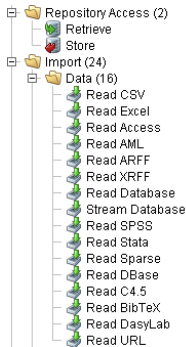
- Violett = ExampleSet
- Grün = Model
- Braun = PerformanceVector
- Pink = Merkmalsgewicht



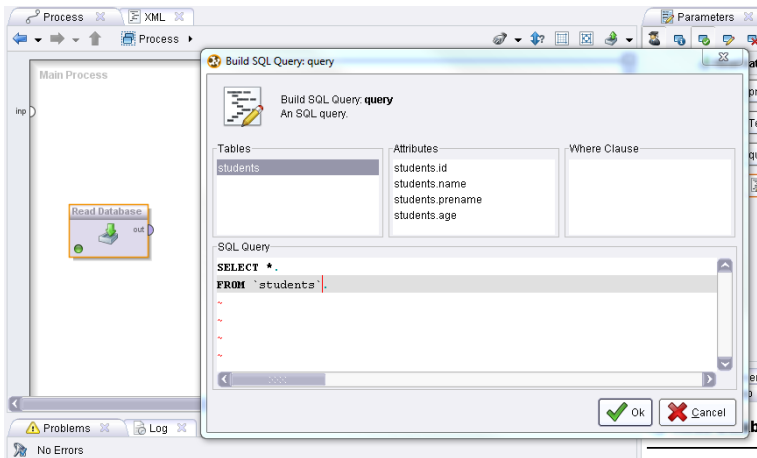


## Datenselektion und -extraktion

- Import aus
  - Repository
  - verschiedenen Dateiformaten
- Generierung von Daten
- Ausgabe als ExampleSet
- Speicherung im Repository möglich



## Lesen aus Datenbank



The screenshot shows the 'Build SQL Query' dialog box in RapidMiner. The main process area on the left contains a 'Read Database' operator. The dialog box is titled 'Build SQL Query: query' and contains the following fields:

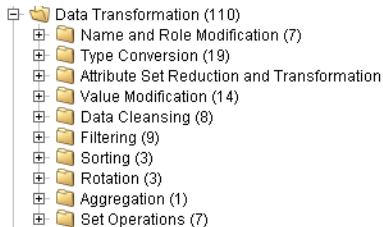
- Tables:** A list box containing 'students'.
- Attributes:** A list box containing 'students.id', 'students.name', 'students.prename', and 'students.age'.
- Where Clause:** An empty text area.
- SQL Query:** A text area containing the SQL query:
 

```
SELECT *.
FROM `students`.
```

At the bottom right of the dialog box are 'Ok' and 'Cancel' buttons. The main window also shows a 'Problems' panel at the bottom left with the message 'No Errors'.

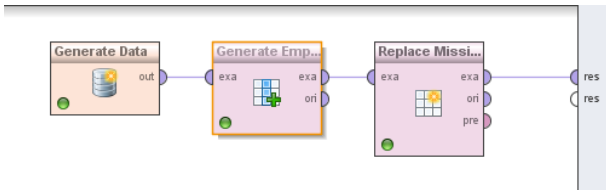
## Datenbereinigung und -transformation

- Umbenennung, Rollenzuweisung
- Typumwandlung
- Transformation von Attributen
- Wertmodifikation
- Datenbereinigung
- Filterung
- Sortierung
- Rotation
- Aggregation
- Operatoren (z.B. Joins)



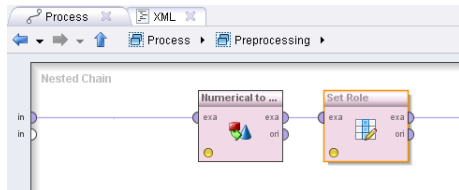
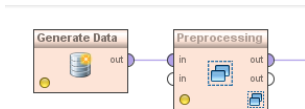
## Beispiel Data Cleansing

- Operator: Replace Missing Values
- Ersetzung fehlender Werte durch
  - Minimum
  - Maximum
  - Durchschnitt
  - Null
  - Wert



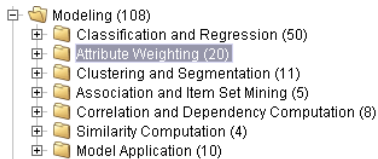
## Vorverarbeitung als Subprozess

- Vorverarbeitungsschritte als Subprozess gekapselt  
→ bessere Übersicht
- Utility/Subprocess



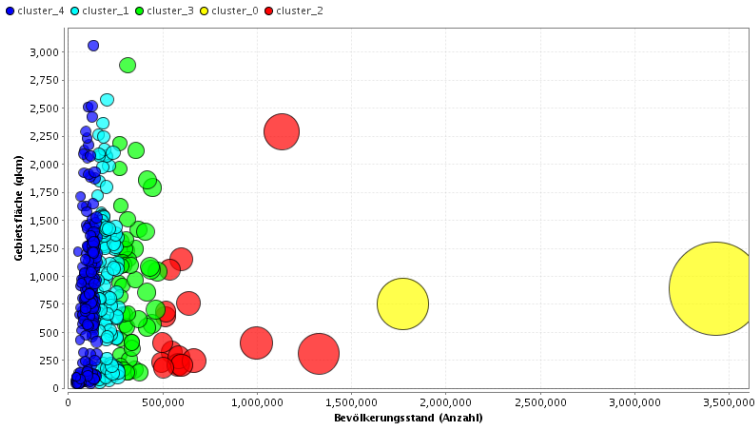
# Data Mining

- Klassifikation
- Attributgewichtung
- Clustering
- Assoziationsanalyse
- Korrelation
- Ähnlichkeitsberechnung



- [-] Modeling (108)
  - [+] Classification and Regression (50)
  - [+] Attribute Weighting (20)
  - [+] Clustering and Segmentation (11)
  - [+] Association and Item Set Mining (5)
  - [+] Correlation and Dependency Computation (8)
  - [+] Similarity Computation (4)
  - [+] Model Application (10)

## Beispiel Clustering



## Interpretation

- visuelle Darstellung in verschiedenen Graphen- und Diagrammtypen
- Bewertung durch Benutzer
  - gefundene Muster beurteilen
  - Aussagekraft des Ergebnisses
- evtl. erneutes Data Mining



## Produktpräsentation



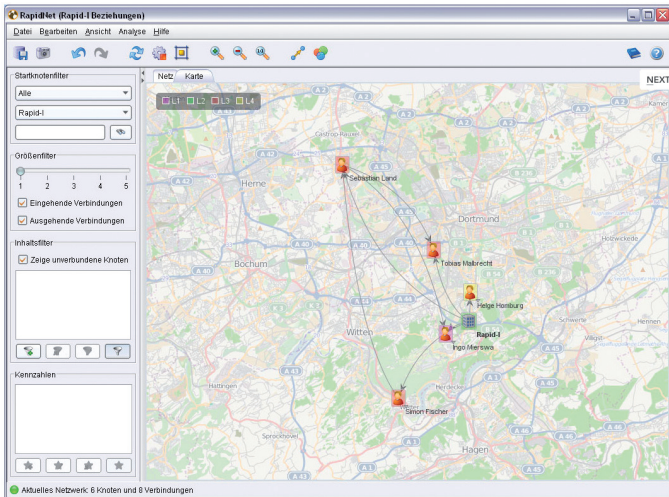
- 1 Allgemeine Übersicht
- 2 Warenkorbanalyse (FP-Growth)
- 3 Clustering nach Einwohnerzahlen
- 4 Text Mining

# Überblick

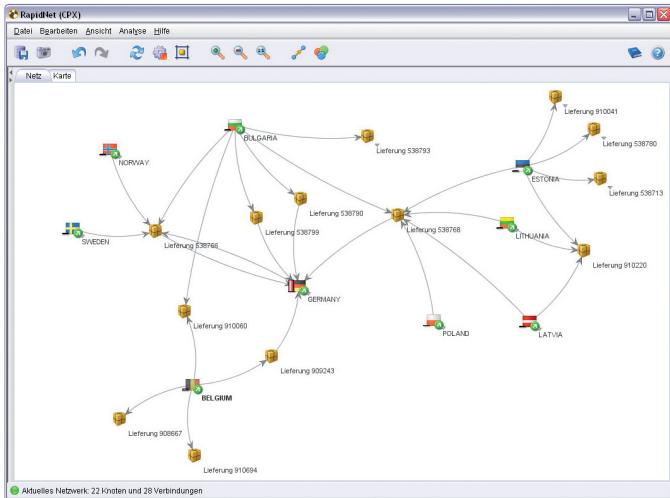
- 1 Das Projekt RapidMiner
- 2 Funktionen
- 3 KDD-Prozess
- 4 Weitere Werkzeuge von Rapid-I**
- 5 Zusammenfassung

## RapidNet

- Allgemein: Struktur- und Relations-Explorer
- zahlreiche Möglichkeiten zur Visualisierung
- basiert auf Funktionen des RapidMiner
- einsatzfähig auf allen gängigen Plattformen
- Möglichkeiten
  - Strukturelle Clusteranalysen
  - Darstellung von hierarchischen Relationen
  - Visualisierung von geographischen Informationen auf Karten
  - ...



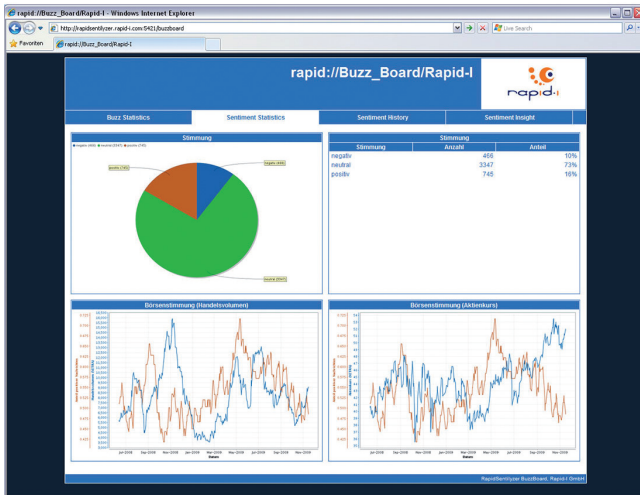
(Quelle: [RI10b])



(Quelle: [RI10b])

## RapidSentryzer

- dient zur automatischen Sammlung von Informationen
- Verwendung von Crawling-Techniken in Kombination mit Data- und Text Mining
- basiert auf Funktionen des RapidMiner
- zentrale Zusammenfassung der Informationen im sogenannten „RapidSentryzer BuzzBoard“, bestehend aus:
  - Buzz Statistics
  - Sentiment Statistics
  - Sentiment History
  - Sentiment Insight



(Quelle: [RI10c])

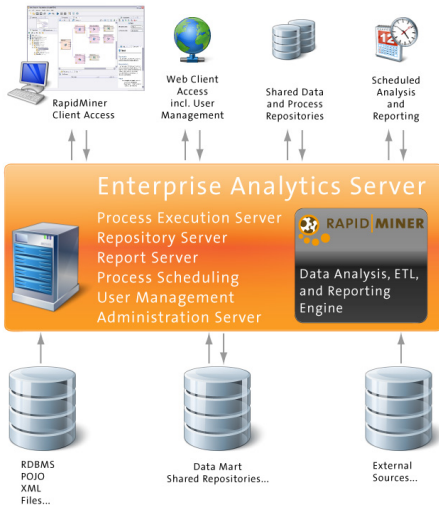


(Quelle: [RI10c])



## RapidAnalytics

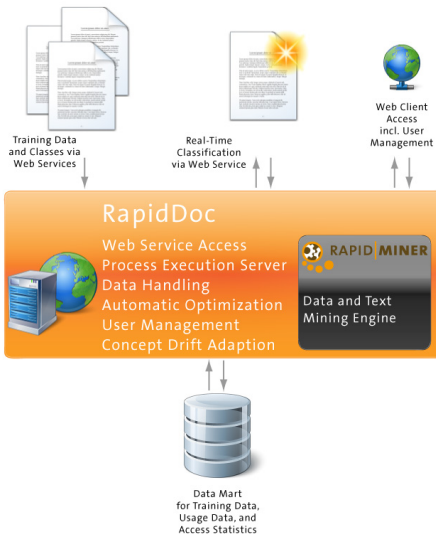
- Open Source Enterprise Analytics Server
- basierend auf RapidMiner
- Shared Repositories
- Remote und Scheduled Execution
- Zugriff über
  - RapidMiner Client Software
  - Web-Interface
  - Webservices



(Quelle: [RIa])

## RapidDoc

- automatische Klassifikation von Texten
- Funktionsweise
  - Basis: Webservices
  - Trainingstexte und vordef. Klassen vorgeben
  - Einordnung neuer Text in wahrscheinlichste Klasse
  - Angabe der Sicherheit der Vorhersage
  - Nutzung der RapidMiner Engine
  - Optimierung durch Rapid-I Mitarbeiter



(Quelle: [R1b])

## Zusammenfassung

- mächtiges Open Source-Data Mining-Tool
- ermöglicht gesamten KDD-Prozess
- viele Operatoren bereits vorhanden
- flexibel einsetz- und erweiterbar
- zahlreiche Visualisierungsvarianten
- weitere Möglichkeiten durch zusätzliche Tools
- Einsatz z.B. bei
  - Allianz
  - Siemens
  - EADS
  - T-Mobile
  - PC-Ware

## Quellen I

- [RIa] RAPID-I: *Rapid Analytics*. [http://rapid-i.com/component/option,com\\_docman/task,doc\\_download/gid,48/](http://rapid-i.com/component/option,com_docman/task,doc_download/gid,48/). – Zugriff: 22.04.2010
- [RIb] RAPID-I: *Rapid Doc*. [http://rapid-i.com/component/option,com\\_docman/task,doc\\_download/gid,49/](http://rapid-i.com/component/option,com_docman/task,doc_download/gid,49/). – Zugriff: 22.04.2010
- [RIc] RAPID-I: *RapidMiner Benutzerhandbuch*. <http://sourceforge.net/projects/yale/files/1.%20RapidMiner/5.0/rapidminer-5.0-manual-german.pdf/download>. – Zugriff: 09.05.2010
- [RI09] RAPID-I: *RapidMiner 4.4*. <http://ignum.dl.sourceforge.net/project/yale/1.%20RapidMiner/4.4/rapidminer-4.4-tutorial.pdf>. Version: März 2009. – Zugriff: 18.04.2010

## Quellen II

- [RI10a] RAPID-I: *Rapid - I - RapidMiner*.  
<http://rapid-i.com/content/view/181/190/>.  
Version: April 2010. – Zugriff: 20.04.2010
- [RI10b] RAPID-I: *RapidNet*. [http://rapid-i.com/component/option,com\\_docman/task,doc\\_download/gid,50/](http://rapid-i.com/component/option,com_docman/task,doc_download/gid,50/).  
Version: Februar 2010. – Zugriff: 23.04.2010
- [RI10c] RAPID-I: *RapidSentlyzer*. [http://rapid-i.com/component/option,com\\_docman/task,doc\\_download/gid,51/](http://rapid-i.com/component/option,com_docman/task,doc_download/gid,51/).  
Version: Februar 2010. – Zugriff: 23.04.2010
- [TU ] TU DORTMUND: *Data Mining mit RapidMiner*.  
<http://www-ai.cs.uni-dortmund.de/LEHRE/VORLESUNGEN/KDD/SS09/RapidMiner.pdf>. – Zugriff:  
22.04.2010
- [Wik10] WIKIPEDIA: *RapidMiner*.  
<http://de.wikipedia.org/wiki/RapidMiner>. Version: April  
2010. – Zugriff: 18.04.2010

Vielen Dank für die Aufmerksamkeit!