Applied Analytical Data Science

Teil 12: Recommender Systeme & Intelligent Discovery Assistants

Dr. Jörg-Uwe Kietz,
Vorlesung an der Univ. Zürich,

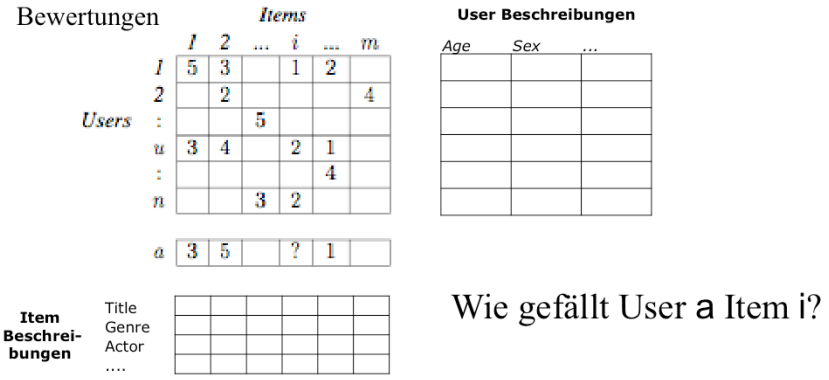
Mittwoch, 14:00-15:45 Uhr Vorlesung,
16:00-17:30 Uhr Übung

http://www.kietz.ch/AADS/

Heutiges Programm

- Recommender Systeme Definition & Anwendungen
  - Basis der Recommendations
    - Collaborative Filtering
    - Content/Demographic-based Recommendation
    - Hybride Ansätze
  - Methoden
    - Instanz-basiert (z.B. k-nearest neighbour)
    - Model-basiert (z.B. clustering)
    - Faktor-basiert (z.B. singular value decomposition (SVD))
  - Evaluation
- Intelligent Discovery Assistants

Recommender Systeme: Problem Definition



Wie gefällt User a Item i?

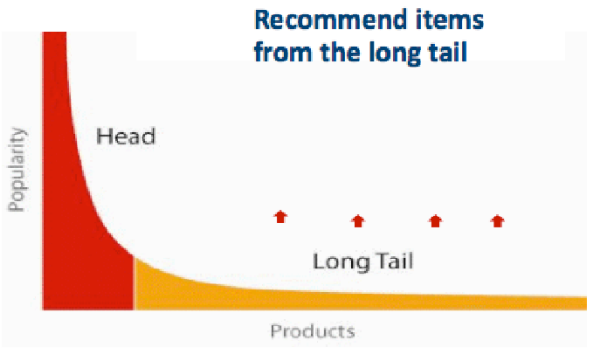
Welche User finden welche Items, die sie noch nicht kennen, gut?

Recommender Systeme Anwendungen

- Netflix (Filmverleih)
- Amazon (Produktverkauf)
- iTunes Genius (DJ, Verkauf)
- LinkedIn (Kontakt & Job Vermittlung)
- Google (Suchmachinenanworten, Welche Werbung)
- …

Personalisierte Empfehlungen

The long Tail & Sparsity



A long tail also mean only a few ratings are available (sparse matrix)

Heutiges Programm

- Recommender Systeme Definition & Anwendungen
- Basis der Recommendations
  - Collaborative Filtering
  - Content/Demographic-based Recommendation
  - Hybride Ansätze
- Methoden
  - Instanz-basiert (z.B. k-nearest neighbour)
  - Model-basiert (z.B. clustering)
  - Faktor-basiert (z.B. singular value decomposition (SVD))
- Evaluation

Collaborative Filtering

Bewertungen

		Items					
		1	*2*	...	*i*	...	*m*
Users	*1*	5	3		1	2	
	2		2				4
	:			5			
	u	3	4		2	1	
	:					4	
	n			3	2		
	a	3	5		?	1	

Content-based Recommendation

Bewertungen

		Items					
		1	*2*	...	*i*	...	*m*
Users	*1*	5	3		1	2	
	2		2				4
	:			5			
	u	3	4		2	1	
	:					4	
	n			3	2		
	a	3	5		?	1	

Item Beschrei-bungen

Title						
Genre						
Actor						
....						

Demographic-based Recommendation

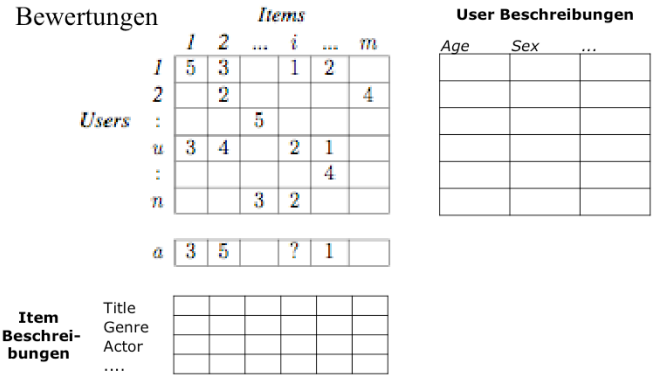
Bewertungen

		Items					
		1	*2*	...	*i*	...	*m*
Users	*1*	5	3		1	2	
	2		2				4
	:			5			
	u	3	4		2	1	
	:					4	
	n			3	2		
	a	3	5		?	1	

User Beschreibungen

Age	*Sex*	...

Hybride Ansätze



Heutiges Programm

- Recommender Systeme Definition & Anwendungen
- Basis der Recommendations
  - Collaborative Filtering
  - Content/Demographic-based Recommendation
  - Hybride Ansätze
- Methoden
  - Instanz-basiert (z.B. k-nearest neighbour)
  - Model-basiert (z.B. clustering)
  - Faktor-basiert (z.B. singular value decomposition (SVD))
- Evaluation

Methoden: Instanz-basiert

- Finde |N| ähnliche
  - Items, und/oder
  - User
- Bestimme das Rating basierend auf den Durchschnittsratings und der Ähnlichkeit:

$$pred(a,p) = \overline{r_a} + \frac{\sum_{b \in N} sim(a,b) * (r_{b,p} - \overline{r_b})}{\sum_{b \in N} sim(a,b)}$$

Ähnlichkeitsmasse für CF

User-basiertes CF: Pearson correlation

$$sim(a,b) = \frac{\sum_{p \in P}(r_{a,p} - \bar{r}_a)(r_{b,p} - \bar{r}_b)}{\sqrt{\sum_{p \in P}(r_{a,p} - \bar{r}_a)^2} \sqrt{\sum_{p \in P}(r_{b,p} - \bar{r}_b)^2}}$$

Item-basiertes CF: Cosine Similarity (Winkel zwischen den Vektoren)

$$sim(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| * |\vec{b}|}$$

Mit normalisierten (durchschnittliche Bewertungen) Vektoren:

$$sim(a,b) = \frac{\sum_{u \in U}(r_{u,a} - \overline{r_u})(r_{u,b} - \overline{r_u})}{\sqrt{\sum_{u \in U}(r_{u,a} - \overline{r_u})^2} \sqrt{\sum_{u \in U}(r_{u,b} - \overline{r_u})^2}}$$

Similarity measures & Sparsity

- Which user is more similar to user-1, 2 or 3?

user-1	7	7	7							1		3				
user-2			7	7	7			3	1							
user-3	5	9	4			1	3								1	

- What should a similarity measure measure?
  - How similar are the overlapping parts
  - How large is the overlapping part
    - With respect to the total of items?
    - With respect to the union of their items?
  - How should each component be weighted?

Model-basiert

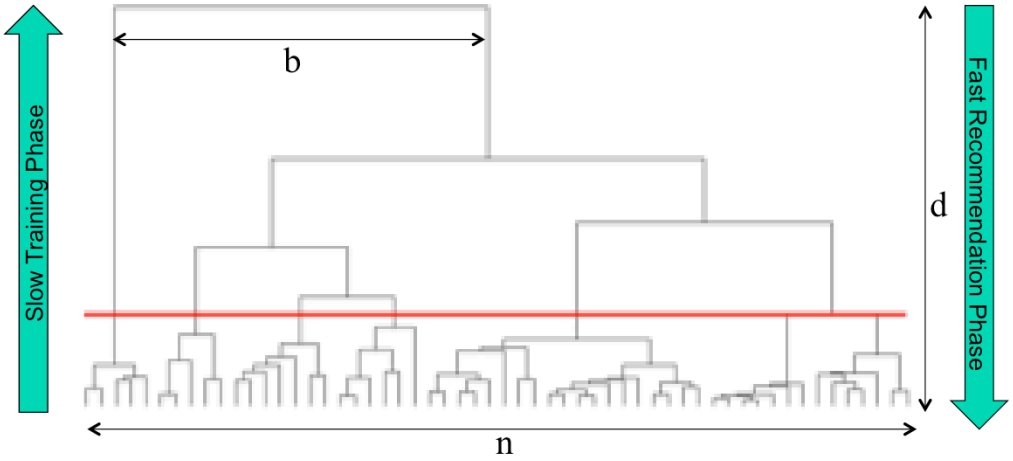
- Lerne ein (Cluster-) Modell aus den Daten
- Benutze das Modell zur Vorhersage

Vorteile:

⇒ Schneller zur Vorhersage (als k-NN)

⇒ Berücksichtigung der Ratings auch in Cold-Start Situationen

Hierachical Clustering for Scalability



Ideally n >> d, i.e. for a balanced clustering $d = \log_b n$

Clustering for Cold-Start

- A new item has not rated by anyone so far

⇒ Only item-description

- If k-NN looks for similar items

⇒ All ratings are ignored in similarity, as the current one has none

If you build a clustering based in item-description & ratings

⇒ Items in a cluster are similar in description & rating

⇒ If a new items is classified into a cluster it “inherits” the ratings

Methoden: Faktor basiert

- Bestimme die (k wichtigsten, mit k = 20 bis 100) Faktoren/ Hauptcomponenten mit denen sich die Rating-Matrix approximieren lässt.

- Problem: Faktoren/Hauptkomponenten analyse funktioniert nicht mit sparsen Matrixen
  => verschiedene Methoden in der Literatur
  => convexe Optimierung der Fehler (http://www.stanford.edu/ ~hastie/TALKS/SVD_hastie.pdf)

SVD zur Matrix Approximaton

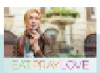
- Finde eine Rank k Approximation:
  $A_{ui} = U_{uk} \times S_{kk} \times (V_{ik})^T$ mit $k << u,i$
  - U_{ur} approx.der User-Ähnlichkeit in k dimensionen
  - V_{ir} approx. der Item-Ähnlichkeit in k dimensionen
  - S_{rr} Gewichtung (Singular-Werte) der k dimensionen

- Für alle benötigten Vektoralgebra Definitionen und ein Beispiel zur Berechnung von U, S und V siehe: Kirk Baker, Singular Value Decomposition Tutorial
  http://www.ling.ohio-state.edu/~kbaker/pubs/Singular_Value_Decomposition_Tutorial.pdf

SVD Beispiel

- **SVD:** $M_k = U_k \times \Sigma_k \times V_k^T$

U_k	Dim1	Dim2
Alice	0.47	-0.30
Bob	-0.44	0.23
Mary	0.70	-0.06
Sue	0.31	0.93

V_k^T					
Dim1	-0.44	-0.57	0.06	0.38	0.57
Dim2	0.58	-0.66	0.26	0.18	-0.36

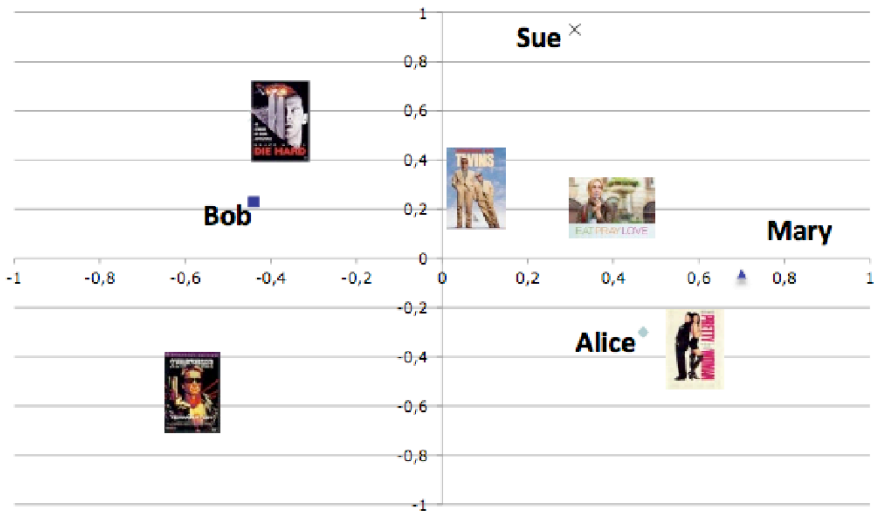
Σ_k	Dim1	Dim2
Dim1	5.63	0
Dim2	0	3.23

- **Prediction:** $\hat{r}_{ui} = \bar{r}_u + U_k(Alice) \times \Sigma_k \times V_k^T(EPL)$

  = 3 + 0.84 = 3.84

Aus: Jannach, Friedrich: Tutorial Recommender Systems, IJCAI, 2013
http://ijcai13.org/files/tutorial_slides/td3.pdf

SVD Beispiel



Heutiges Programm

- Recommender Systeme Definition & Anwendungen
- Basis der Recommendations
  - Collaborative Filtering
  - Content/Demographic-based Recommendation
  - Hybride Ansätze
- Methoden
  - Instanz-basiert (z.B. k-nearest neighbour)
  - Model-basiert (z.B. clustering)
  - Faktor-basiert (z.B. singular value decomposition (SVD))
- Evaluation

Precision & Recall (IR) Based Evaluation

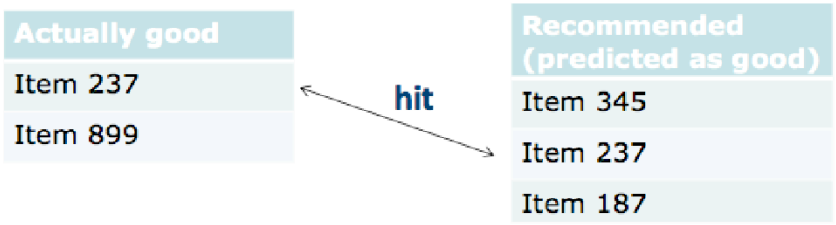
		Reality	
		Actually Good	Actually Bad
Prediction	Rated Good	True Positive (tp)	False Positive (fp)
	Rated Bad	False Negative (fn)	True Negative (tn)

$$Precision = \frac{tp}{tp + fp} = \frac{|good\ movies\ recommended|}{|\text{all recommendations}|}$$

$$Recall = \frac{tp}{tp + fn} = \frac{|good\ movies\ recommended|}{|all\ good\ movies|}$$

Rank Based Score Evaluation

- Pro user:
  Anordnung nach Bewertung Anordnung nach Vorhersage



- Summe aller Ränge/Anzahl der Items*User
- Baseline: Raten = Items/2

Accuracy Measures

- Berechnung des Abstands zwischen wirklicher und vorhergesagter Wertung
  - Mean Absolute Error (*MAE*)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|p_i - r_i|$$

  - Root Mean Square Error (*RMSE*)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i-1}^{n}(p_i - r_i)^2}$$

Fazit

- Recommender Systeme dienen der Personalisierten Empfehlung von Produkten.
- Collaborative Filtering erzeugt Vorhersagen darüber ob jemand etwas mag, basierend auf den Ratings sehr vieler Benutzer.
- Ausser den Ratings kann auch (zusätzlich) Produktinformations und Wissen über die Benutzer genutzt werden.
- Die wichtigsten Methoden sind instanzenbasiert, Model- (clustering-) basiert oder Faktor-basiert.
- Es gibt – je nach Businesszielen - verschiedene Evaluationsmöglichkeiten.

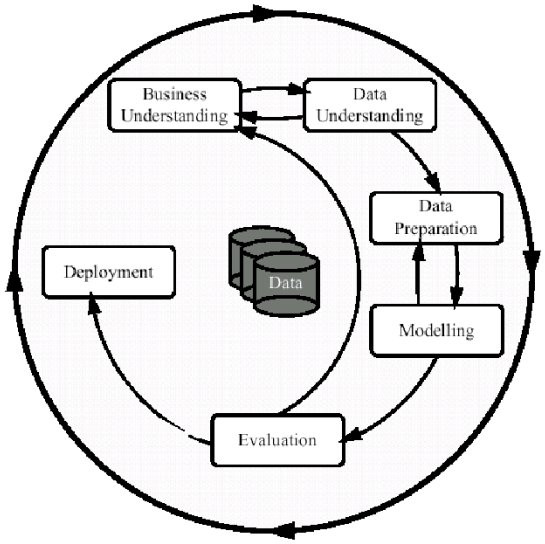
Literatur

- Francesco Ricci, Lior Rokach, Bracha Shapira, Paul B. Kantor (Eds.): Recommender Systems Handbook, Springer, 2011.
- Dietmar Jannach, Gerhard Friedrich: Tutorial: Recommender Systems, International Joint Conference on Artificial Intelligence Beijing, August 4, 2013 (http://ijcai13.org/files/tutorial_slides/td3.pdf)
- Serdar Sali: Using SVD to Predict Movie Ratings (http://classes.soe.ucsc.edu/cmps242/Winter08/proj/serdar_talk.pdf)
- Kirk Baker: Singular Value Decomposition Tutorial (http://www.ling.ohio-state.edu/~kbaker/pubs/Singular_Value_Decomposition_Tutorial.pdf)

Heutiges Programm

- Recommender Systeme Definition & Anwendungen
- Intelligent Discovery Assistants (IDA)
  - What can be supported by an IDA?
  - Who can be supported?
  - How can this be done?

KDD-Process and IDA Support



IDA Support

- What is supported
  - Single step vs. multi-step
  - Graphical editing vs. automatic generation
  - Reuse experience vs. generation from scratch
  - Task decomposition vs. plain plan
  - Design support vs. explanations

- Who is supported:
  - novice vs. expert

What: Single step vs. multi-step

- Most IDAs focus on single step
  Help the user to
  - choose the right (modeling) operator
  - set operator's parameters

- Multi-step
  Advice through all steps of a DM task
  What's the right sequence of operators to solve a DM task

What: Graphical editing vs. automatic generation

- Graphical workflow editing
  - Enable users to drop operators and connect them
  - For large workflows is annoying and time consuming
- Automatic generation
  - Generates all workflows
  - Can find new combinations of operators

What: Reuse experience vs. generation from scratch

- Case based reasoning
  - Store past workflows
  - Learn from them based on goal and data description
- Generate from scratch
  - Redundant workflows
  - No reuse

What: Task decomposition vs. plain plan

- Task decomposition
  - Helps structuring the operators
  - CITRUS and CRISP-DM – decompose DM in a set of tasks with several subtasks
- Plain plan No hierarchy

What: Design support vs. explanations

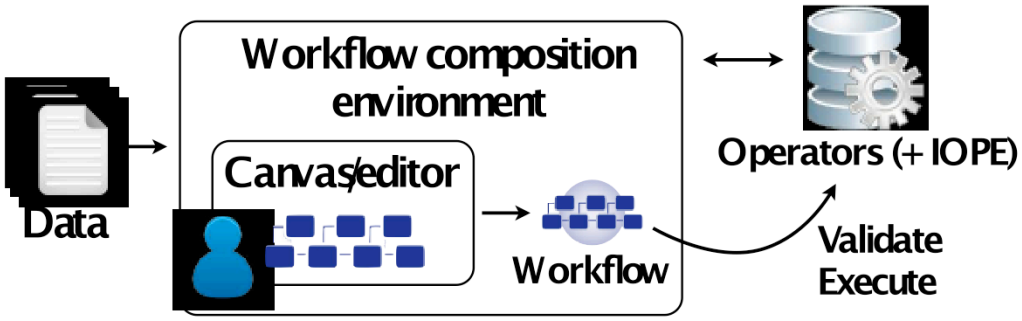
- Design support
  - Help and guidance in how to design a workflow
  - Most of DS have it
- Explanations
  - Why to choose a certain operator
  - Interprets intermediate and final results
  - Describe used techniques and have references to the literature

Who is supported: novice vs. expert

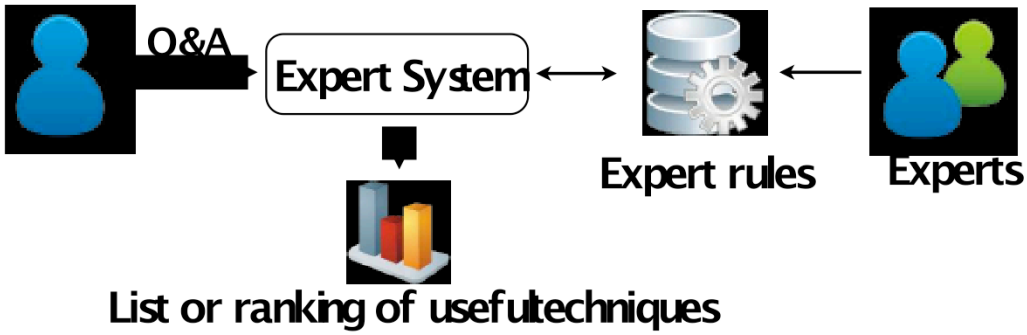
- Which level is support/required
  - novice data miners (often domain experts)
  - expert data miners (often domain novices)

- Support for Novices:
  - How to analyze data correctly?
- Support for expert:
  - How to speed-up design for routine-tasks

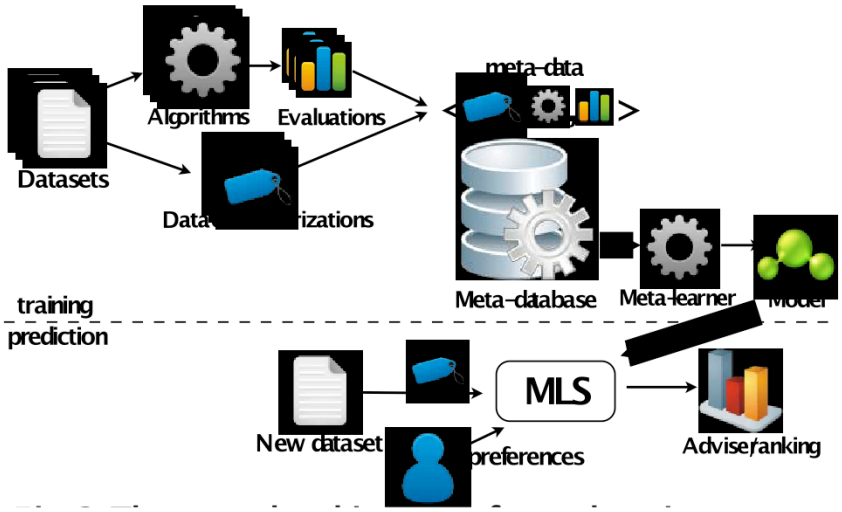
How: Data Mining Support System



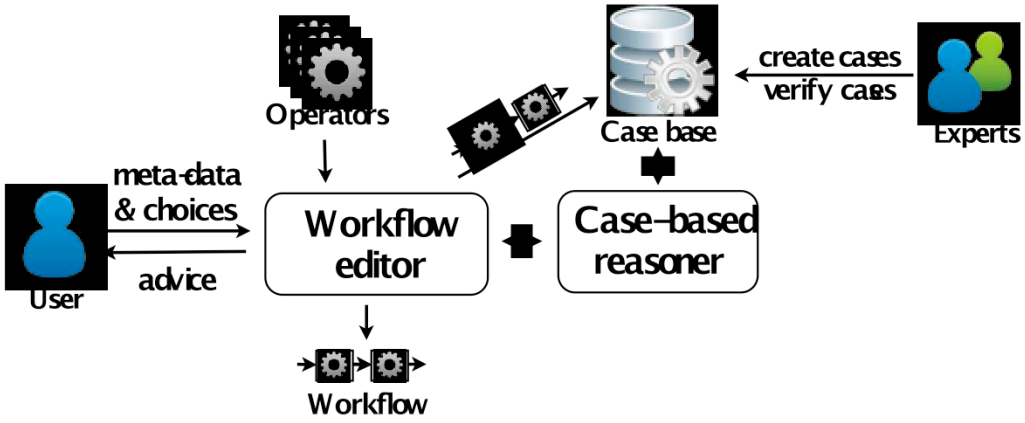
How: Expert System IDA



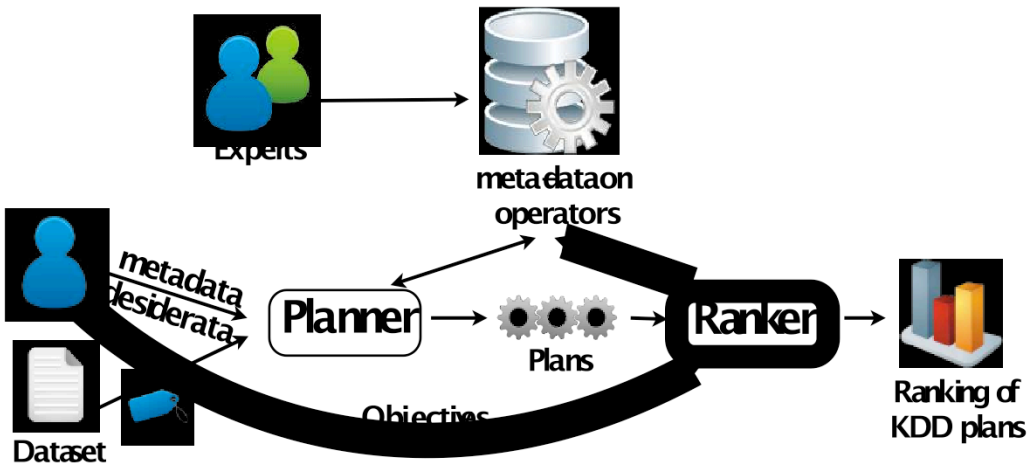
How: Meta-Learning IDA



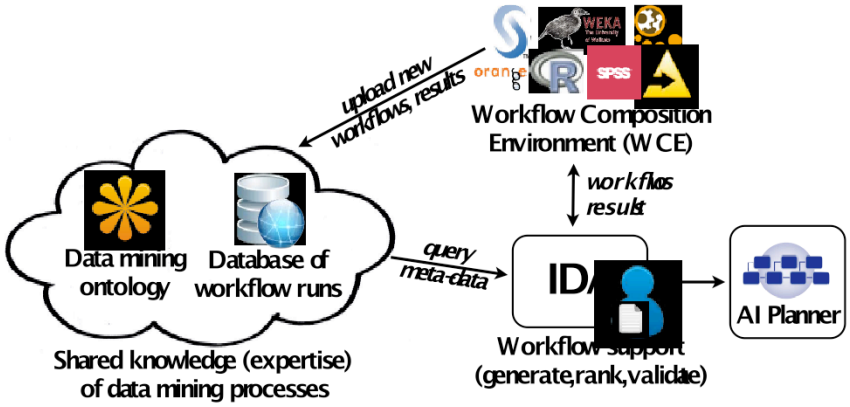
How: Case-based reasoning IDA



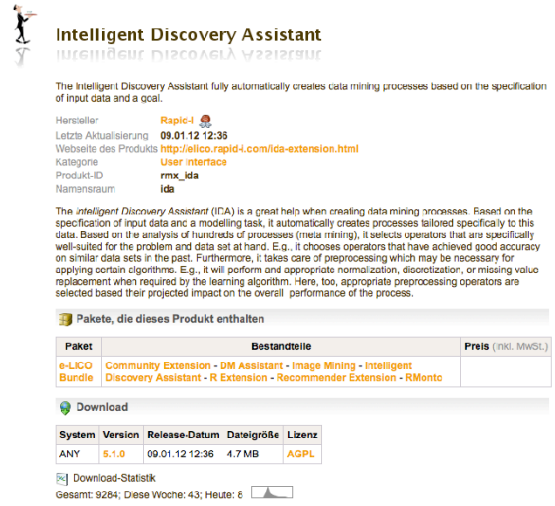
How: AI-Planning IDA



Integration



Our e-Lico IDA



Intelligent Discovery Assistant

The Intelligent Discovery Assistant fully automatically creates data mining processes based on the specification of input data and a goal.

Hersteller	Rapid-I
Letzte Aktualisierung	09.01.12 12:35
Webseite des Produkts	http://elico.rapid-i.com/ida-extension.html
Kategorie	User interface
Produkt-ID	rmx_ida
Namensraum	ida

The *Intelligent Discovery Assistant* (IDA) is a great help when creating data mining processes. Based on the specification of input data and a modelling task, it automatically creates processes tailored specifically to this data. Based on the analysis of hundreds of processes (meta mining), it selects operators that are specifically well-suited for the problem and data set at hand. E.g., it chooses operators that have achieved good accuracy on similar data sets in the past. Furthermore, it takes care of preprocessing which may be necessary for applying certain algorithms. E.g., it will perform and appropriate normalization, discretization, or missing value replacement when required by the learning algorithm. Here, too, appropriate preprocessing operators are selected based their projected impact on the overall performance of the process.

Pakete, die dieses Produkt enthalten

Paket	Bestandteile	Preis (inkl. MwSt.)
e-LICO Bundle	Community Extension - DM Assistant - Image Mining - Intelligent Discovery Assistant - R Extension - Recommender Extension - RMonto	

Download

System	Version	Release-Datum	Dateigröße	Lizenz
ANY	5.1.0	09.01.12 12:36	4.7 MB	AGPL

Download-Statistik

Gesamt: 9284; Diese Woche: 43; Heute: 8

Info:
http://www.e-lico.eu/ida-extension.html

Download:
http://marketplace.rapid-i.com/UpdateServer/faces/product_details.xhtml?productId=rmx_ida

Fazit

- Intelligent Discovery Assistants should help
  - Novices to do data mining correctly
  - Expert to do data mining faster
- There are several methods to build (partial) IDAs,
  - Expert-Systems to use operators correctly
  - Meta-Learning to choose best operators
  - Case-Based Reasoning to use past experiences
  - AI-Planning to build whole DM-workflows
- Existing system can really help
- Integration of (more) different methods would help better

Literatur

Floarea Serban, Joaquin Vanschoren, Jörg-Uwe Kietz, Abraham Bernstein: A survey of intelligent assistants for data analysis, ACM Computing Surveys, 2013
http://dx.doi.org/10.5167/uzh-73010

PDF: http://www.zora.uzh.ch/73010/1/20130201132111_merlin%2Did_6753.pdf