Datenvorverarbeitung von nominalen Daten für Data Mining

Entstanden 2004/2005 bei der
T-Systems International GmbH
unter Betreuung von

Prof. Dr. J. Fürnkranz

Von Oliver Werth

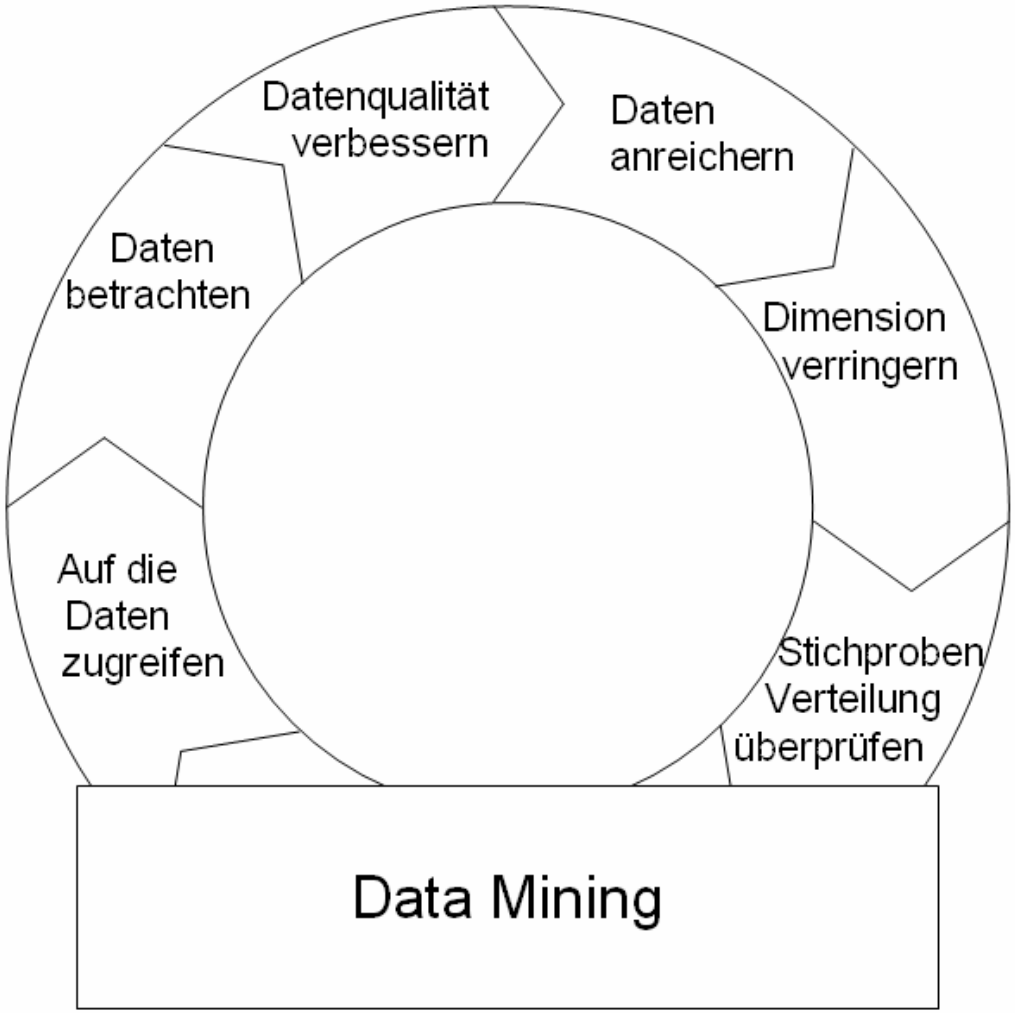
Gliederung

- Datenvorverarbeitung
- Prepared Information Enviroment
- Vorverarbeitung von nominalen Daten
- Exponieren semantischer Information

Datenvorverarbeitung

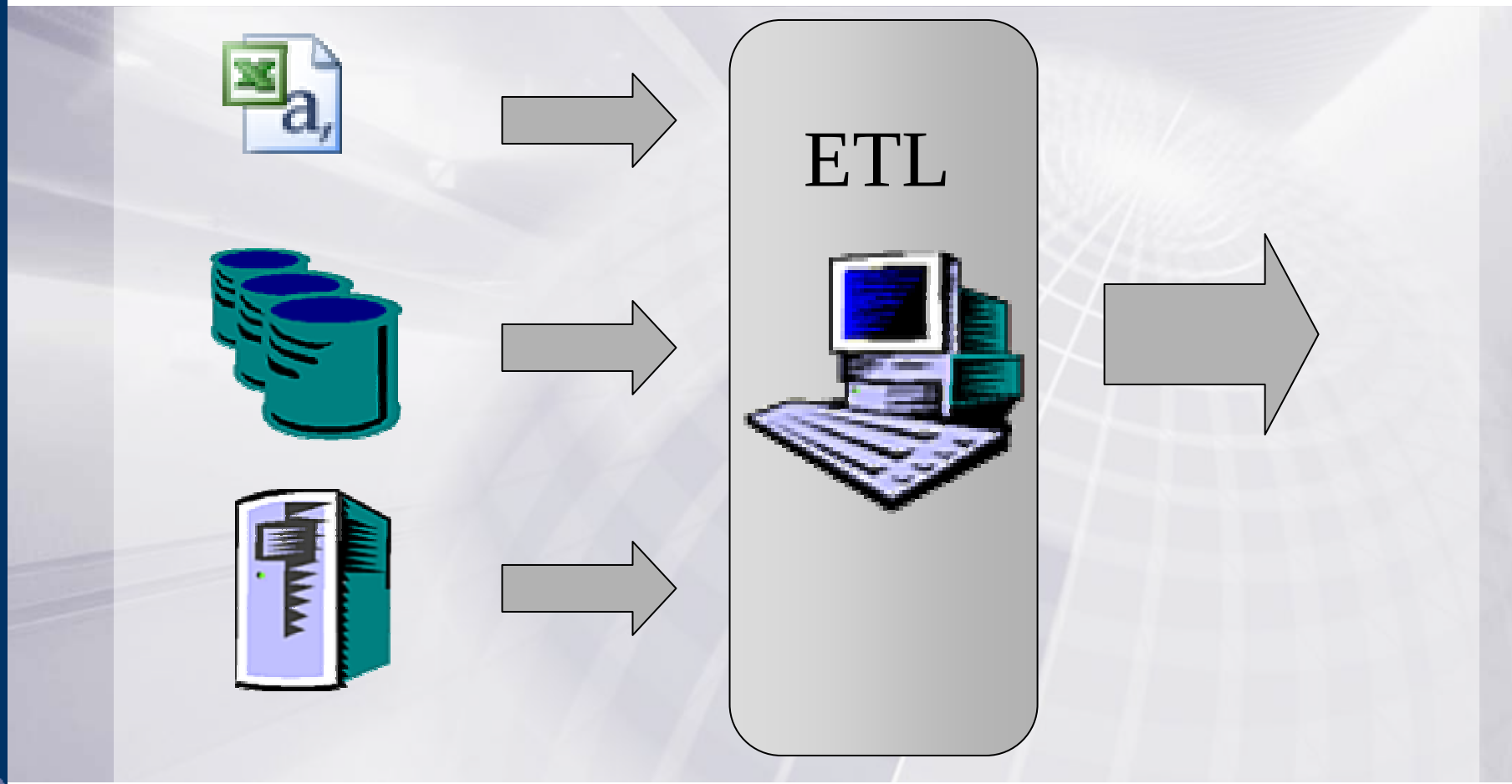
Datenvorverarbeitung

Der Zyklus



Datenvorverarbeitung

Auf die Daten zugreifen



Datenvorverarbeitung
Daten betrachten

- Attribute verstehen
- Fehler erkennen
- Erste Hypothesen

Umgesetzt mit

- OLAP Tools
- SQL

Datenvorverarbeitung
Datenqualität verbessern

- Behandlung von Ausreißern
- Ungültige Werte
- Falsche Formate
- Dubletten
- Fehlende Werte
- Widersprüchliche Werte

Daten anreichern

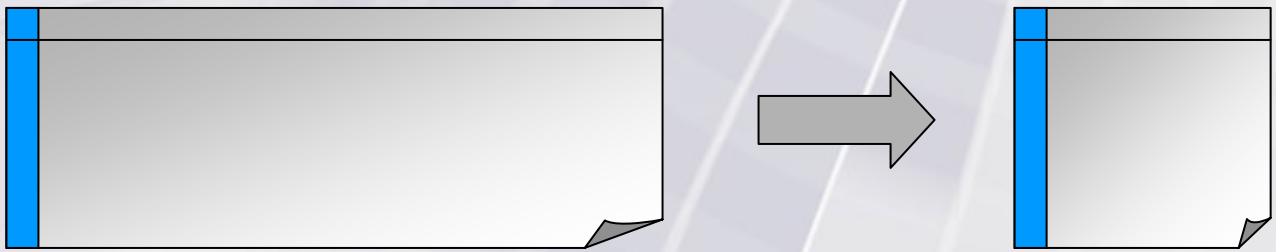
Ergänzen um weitere Attribute

- Aggregationen von Werten
- Attribute aus externen Quellen
- Einbeziehen von Expertenwissen

Domäne des menschlichen Experten

Datenvorverarbeitung
Dimension verringern

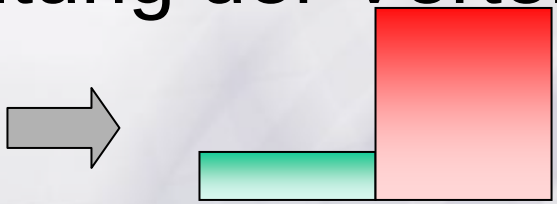
- Vermeiden von Overfitting
- Reduzieren der Rechenzeit
- Entfernen von Selbstbezügen
- Entfernen von Attributen, die in der Praxis nicht zur Verfügung stehen



Datenvorverarbeitung
Stichprobenverteilung prüfen

- Zufällige Verteilung des Targets

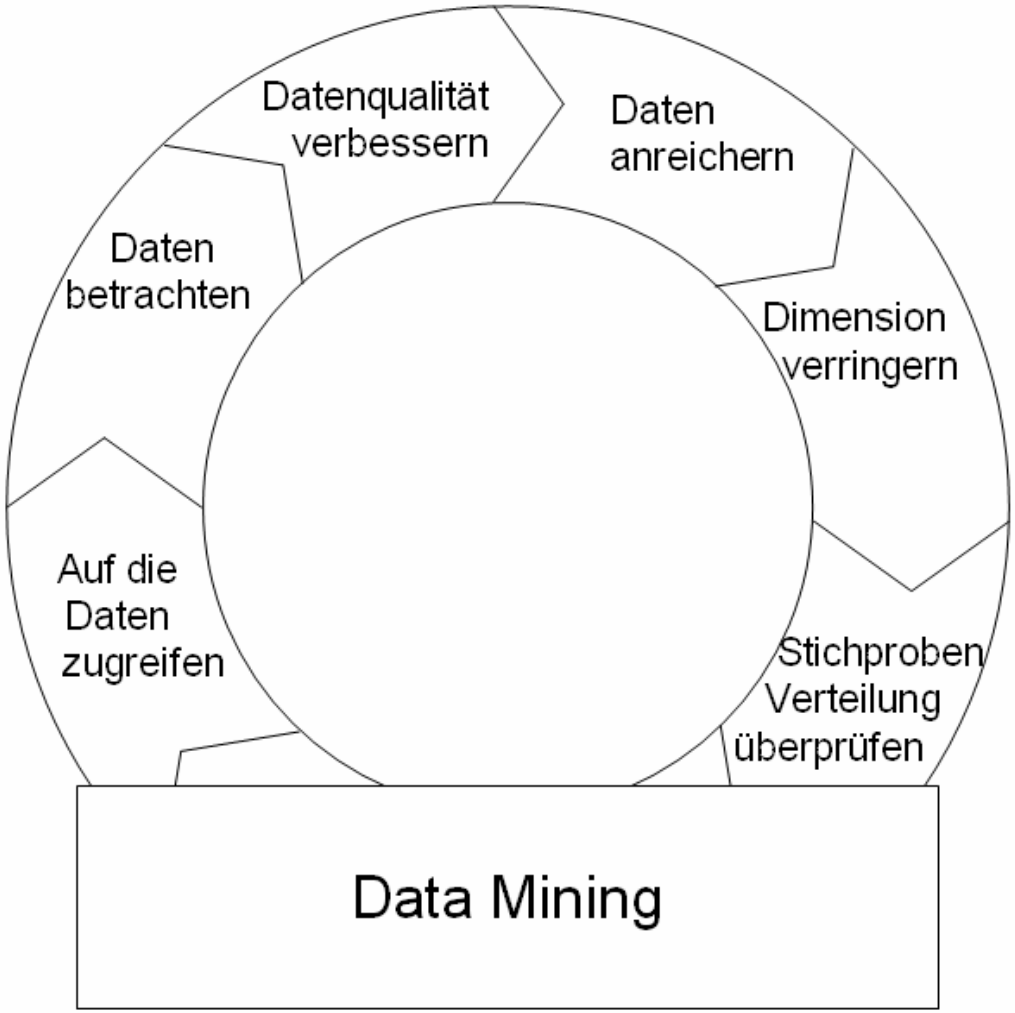
- Erhaltung der Verteilung des Targets



- Gleichmässige Verteilung von Target

Datenvorverarbeitung

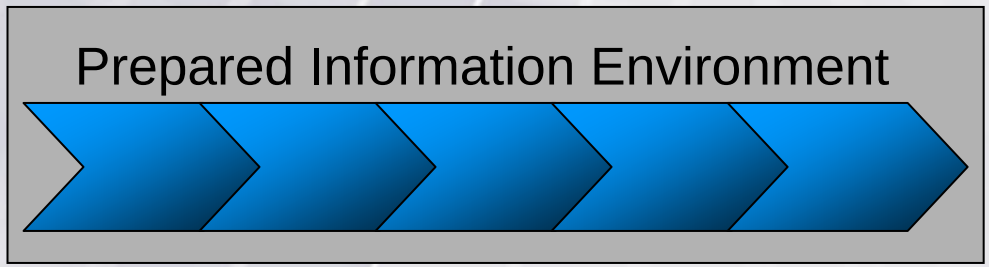
Der Zyklus



Prepared Information Environment

Prepared Information Environment
Überblick

- Ein Rahmen für die Vorverarbeitung
- Vereinfacht das erneute Anstoßen des Zyklus
- Sichert die Replizierbarkeit des Modells
- Routine bei periodisch zu erstellenden Modellen
- Spart Zeit



Prepared Information Environment

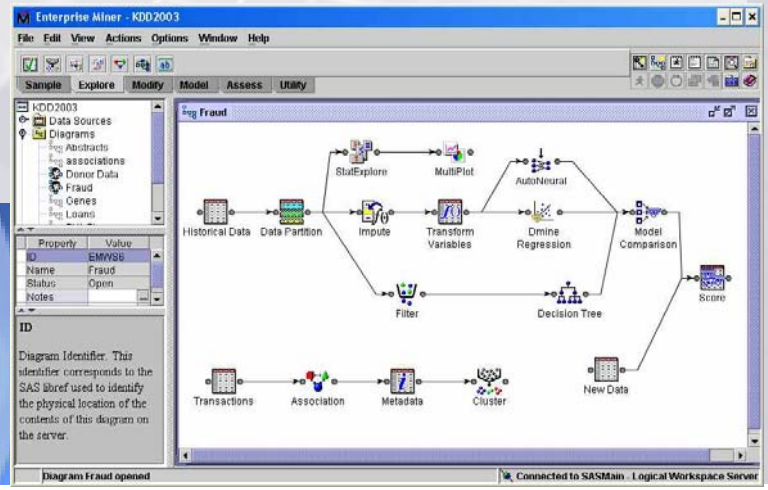
PIE mit SQL

- Transformationen lassen sich mit SQL durchführen
- SQL Skripte können automatisiert erneut aufgerufen werden
- Ist bereits ein Nebenprodukt des Datenvorverarbeitungsprozesses

Prepared Information Environment

PIE mit SAS

- Prozesse in einem Flussdiagramm.
- SAS 9 ist konzipiert um alle Aufgaben der Datenvorverarbeitung durchzuführen
- Bedingt transportabel



Prepared Information Environment

Hybrides PIE

- Zusammenfügen von SQL-Skripten und SAS 8.2
- Output über ID anstelle eines Output-Moduls
- Flexibel
- Geringer Mehraufwand

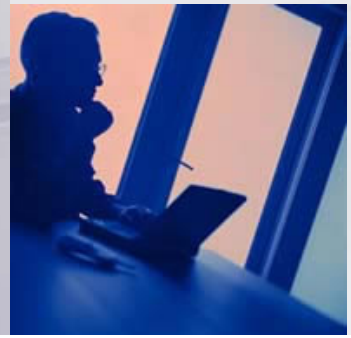
Normalisieren von Schreibweisen

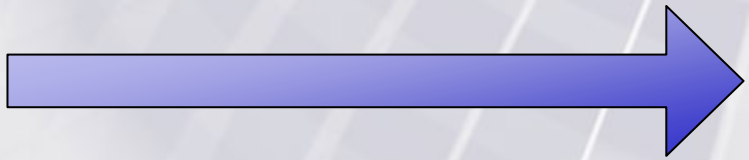
Wissen in der Datenvorverarbeitung

- Unstrukturierte Felder enthalten menschenlesbare Informationen
  - Produktnamen, Titel, Kommentare, Namen,...
- Anreichern der Daten ist eine kreative Arbeit
- Erfahrung und Fachwissen wird für viele Entscheidungen benötigt

Datenbanken

- Expertenwissen zum befüllen notwendig
- Auflistung existenter Nachnamen
  - Befüllung sehr aufwändig
  - Datenbank wird sehr gross
- Angereichert mit Metadaten





Heuristiken

- Levenshtein Distanz
- Präfix Matching
- Jaro Ähnlichkeitsmaß

$$Jaro(s;t) = \frac{1}{3} \cdot \left(\frac{|s'|}{|s|} + \frac{|t'|}{|t|} + \frac{|s'| - T_{s',t'}}{|s'|} \right)$$

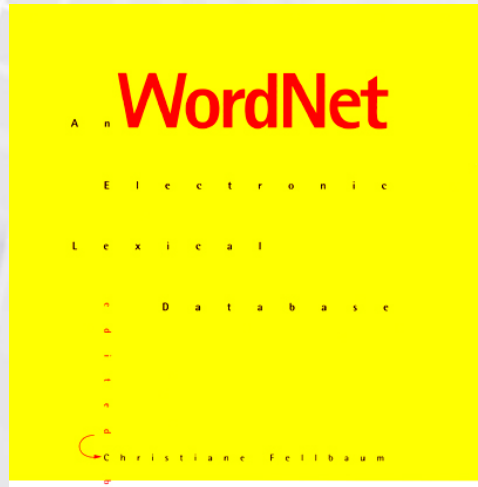
- Jaro-Winkler Maß

$$Jaro\text{-}Winkler(s;t) = Jaro(s;t) + \frac{P'}{10} \cdot (1 - Jaro(s;t))$$

- Smith-Waterman Algorithmus

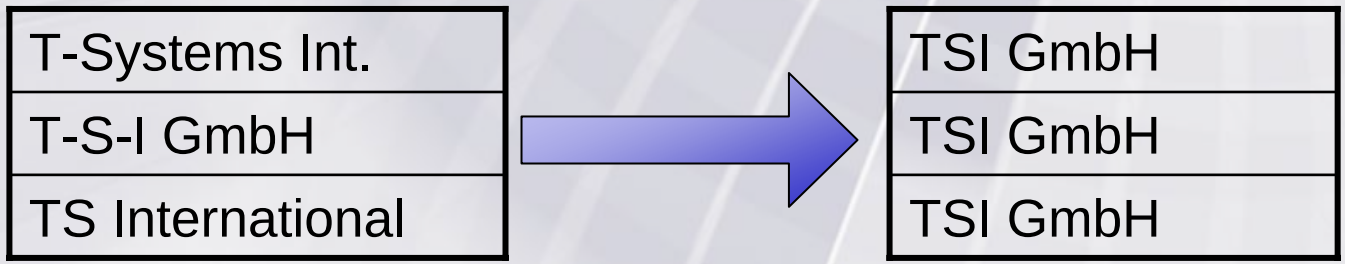
Semantische Netzwerke

- Knoten beinhalten Begriffe
- Kanten repräsentieren Relationen
- Semantische Distanz zwischen Knoten
- Geeignet um Zusammenhänge und Assoziationen zu verarbeiten
  - Homonyme
  - Synonyme
  - Antonyme
  - Hyperonyme
  - Hyponyme
  - Meronyme



Regeldatenbanken

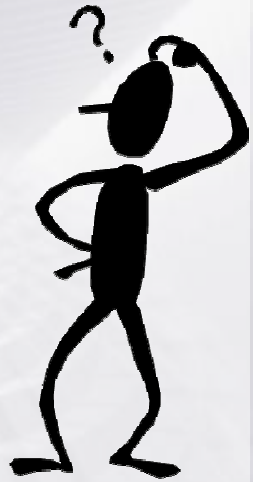
- Regeln können wiederkehrende Fehler beheben
- Regeln können Wortformen normalisieren
- Eine Regeldatenbank enthält Regeln und Referenzen



Normalisieren von Schreibweisen

Problemstellung

- Welche Zeichen sind falsch?
- Welche Zeichen fehlen?
- Welche Zeichen sind verdreht?
- Welche Worte sind falsch?
- Welche Worte sind korrekt, gehören aber woanders hin?
- Welche Worte können unterschiedlich geschrieben werden?



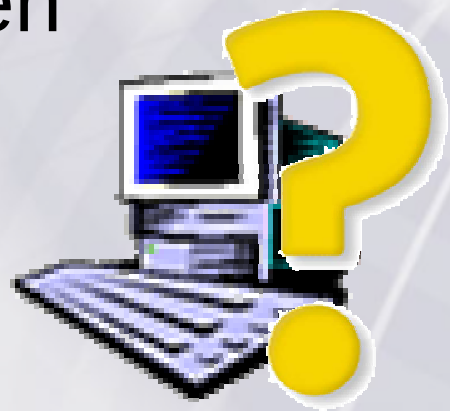
Normalisieren von Schreibweisen

Lösungskonzept

- Eindeutige Form notwendig
- Anwendung von Regeln
- In unbekannten Fällen werden mit Hilfe des Experten weitere Regeln erstellt
- Datenbank mit korrekten Formen
- Vergleich von Formen über Heuristiken

Erkennen von semantischen Fehlern
Problemstellung

- Sind automatisiert fast nicht zu erkennen
- Der Computer kann die menschen-lesbaren Informationen nicht ohne Hilfe verarbeiten



Erkennen von semantischen Fehlern
Lösungskonzept

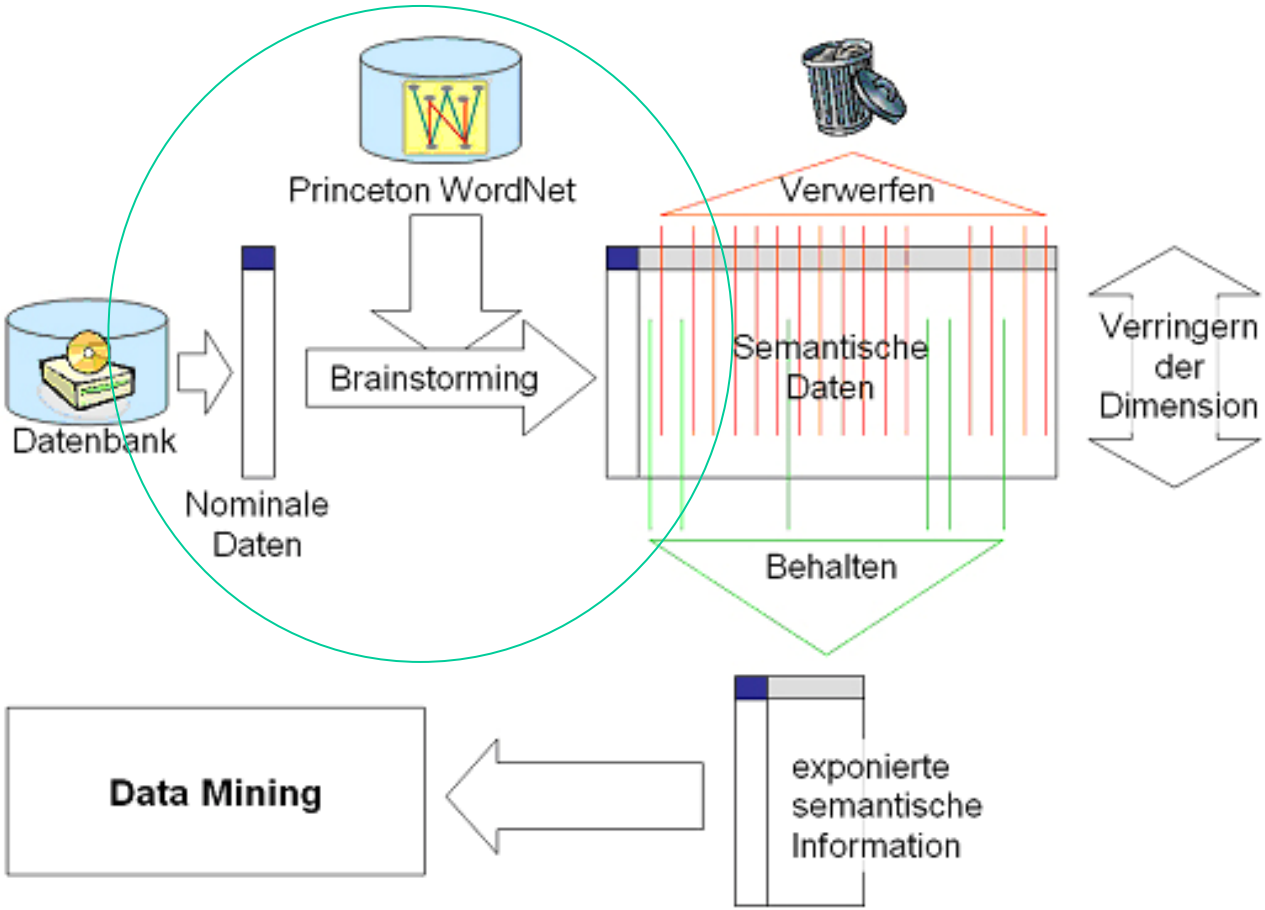
- Sichtung durch einen Experten
- Erkennen des Fehlers am Ende eines Datenvorverarbeitungszyklus
- Zur Automatisierung ist Wissen erforderlich

- Externe Informationen müssten automatisch herangezogen werden.

Exponieren semantischer Information

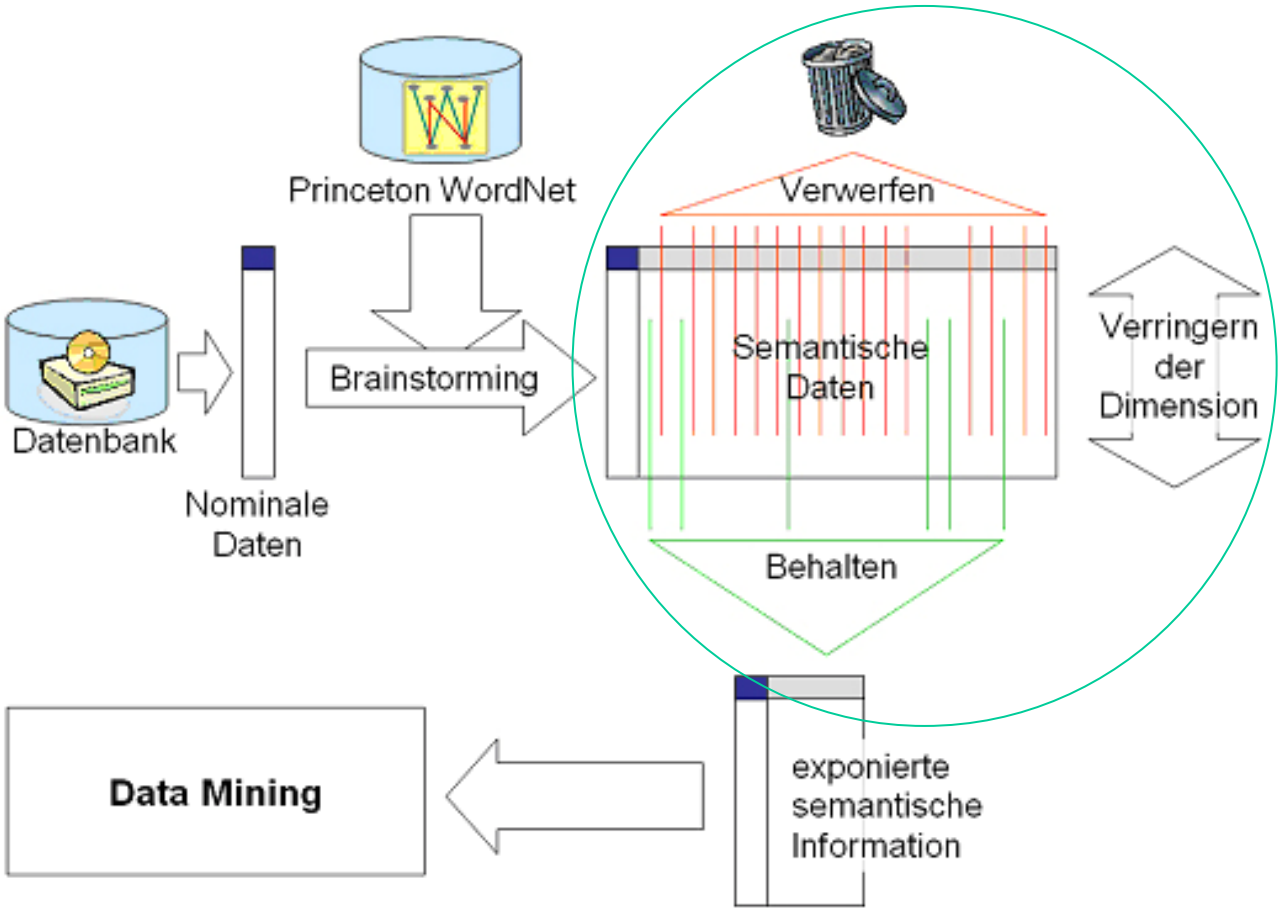
Exponieren semantischer Information

Brainstorming



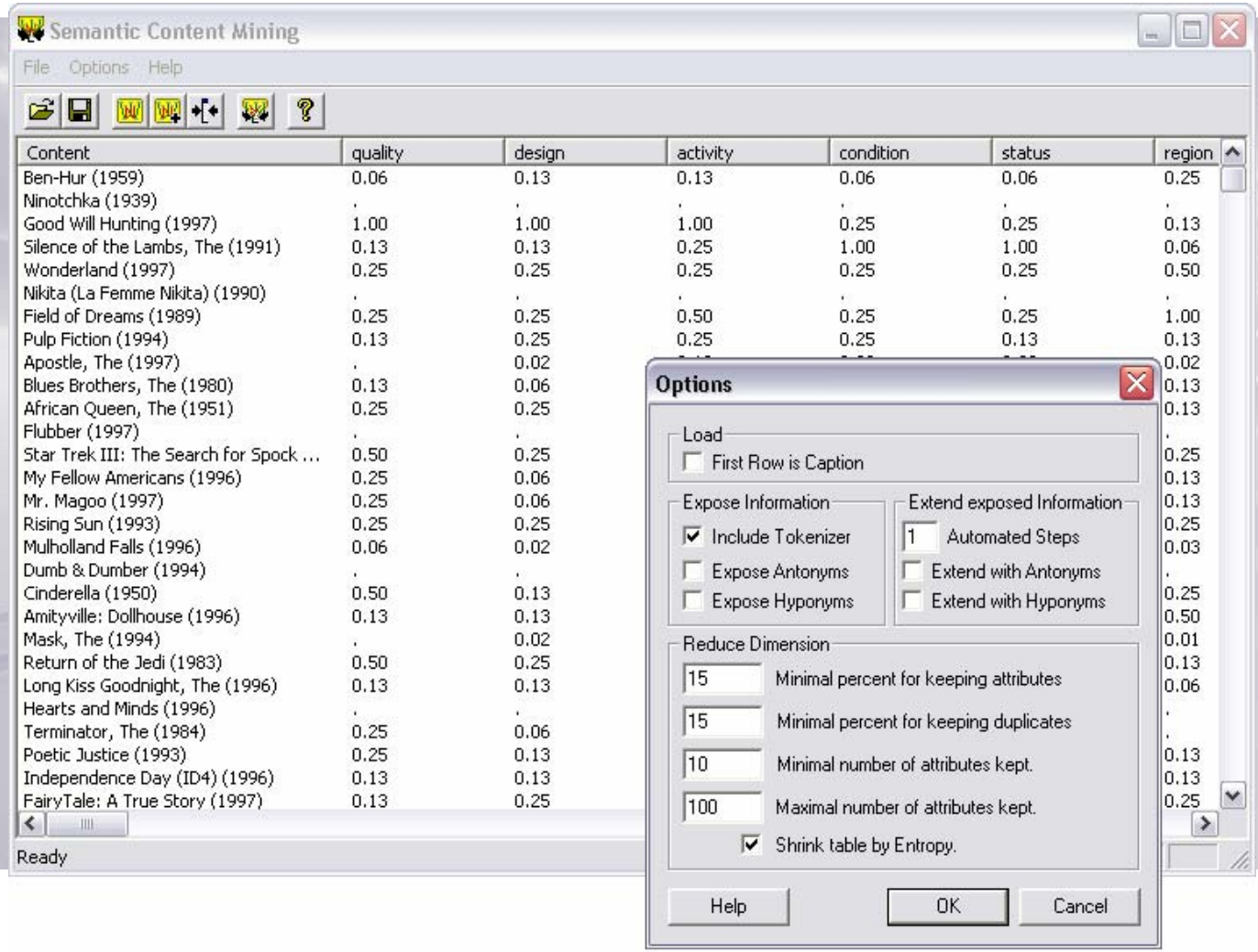
Exponieren semantischer Information

Verringern der Dimension



Exponieren semantischer Information

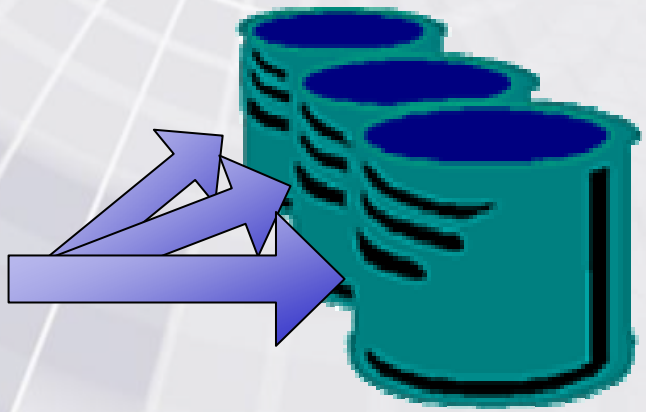
SCM Tool



Evaluationszenario

- Verwendung der offiziellen MovieLens Datenbank
- Prognose der Bewertung anhand des Kinotitels
- Angereichert um manuell erstellte Genre
- Angereichert um mit SCM erstellte Attribute

Up in Smoke (1978)
Two Deaths (1995)
Safe Passage (1994)
Nine Months (1995)
Money Train (1995)
...



Exponieren semantischer Information

Evaluationsresultat

Daten	Wahr-positiv	Gesamter Fehler	Normierte Präzision
Genre	0,0%	21,3%	-1,00
SCM 2.0	8,8%	22,3%	-0,82
SCM 2.5	14,7%	22,9%	-0,70
Genre_SCM 2.0	8,8%	23,9%	-0,82
Genre_SCM 2.5	5,8%	21,3%	-0,88

Vielen Dank für ihre Aufmerksamkeit

Oliver.Werth@Athistaur.de