Automatisierte Dossier-Erstellung mittels Text-Mining

Paul Assendorp

Grundseminar

Gliederung

- Motivation
- Textmining
- Tools
- Aktueller Stand
- Ausblick
- Konferenzen & Forschung

Gliederung

- Motivation
- Textmining
- Tools
- Aktueller Stand
- Ausblick
- Konferenzen & Forschung

Motivation

„Editors don‘t scale“

(nach Dr. Carsten Brosda, Amt Medien, Senatskanzlei Hamburg)

Was ist ein Dossier?

- Zusammenstellung von Dokumenten
- Zu thematischem Hintergrund
- politisch, historisch oder kulturell

Aber: Keine eindeutige Definition

Automatisierte Dossier-Erstellung

"We emphasize that the complexity of language implies that automated content analysis methods […] are best thought of as amplifying and augmenting careful reading and thoughtful analysis"

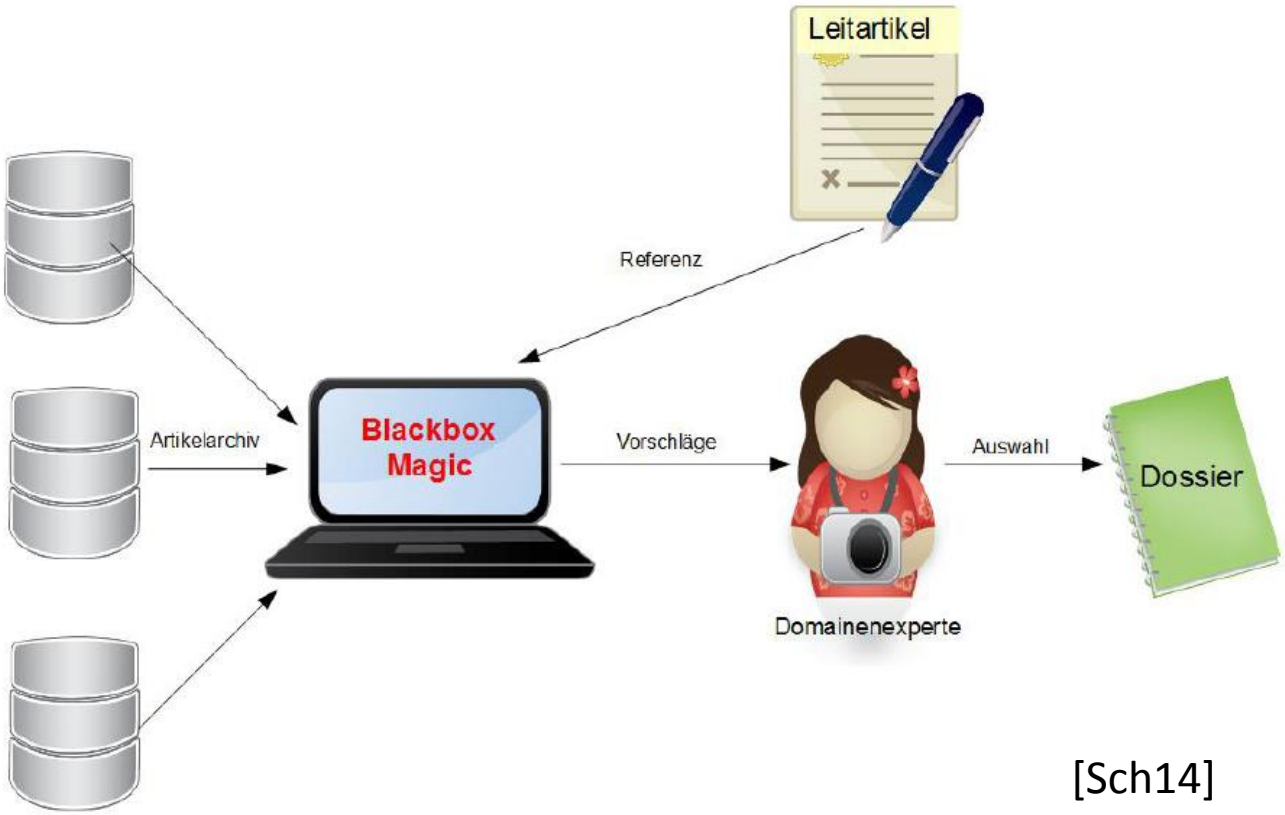
[Grim13]

Automatisierte Dossier-Erstellung (2)

Ansatz:

- Optimierung der Dossier-Erstellung
- Vorschläge für Domäne-Experten
- Leitartikel als Vorgabe

Automatisierte Dossier-Erstellung (3)

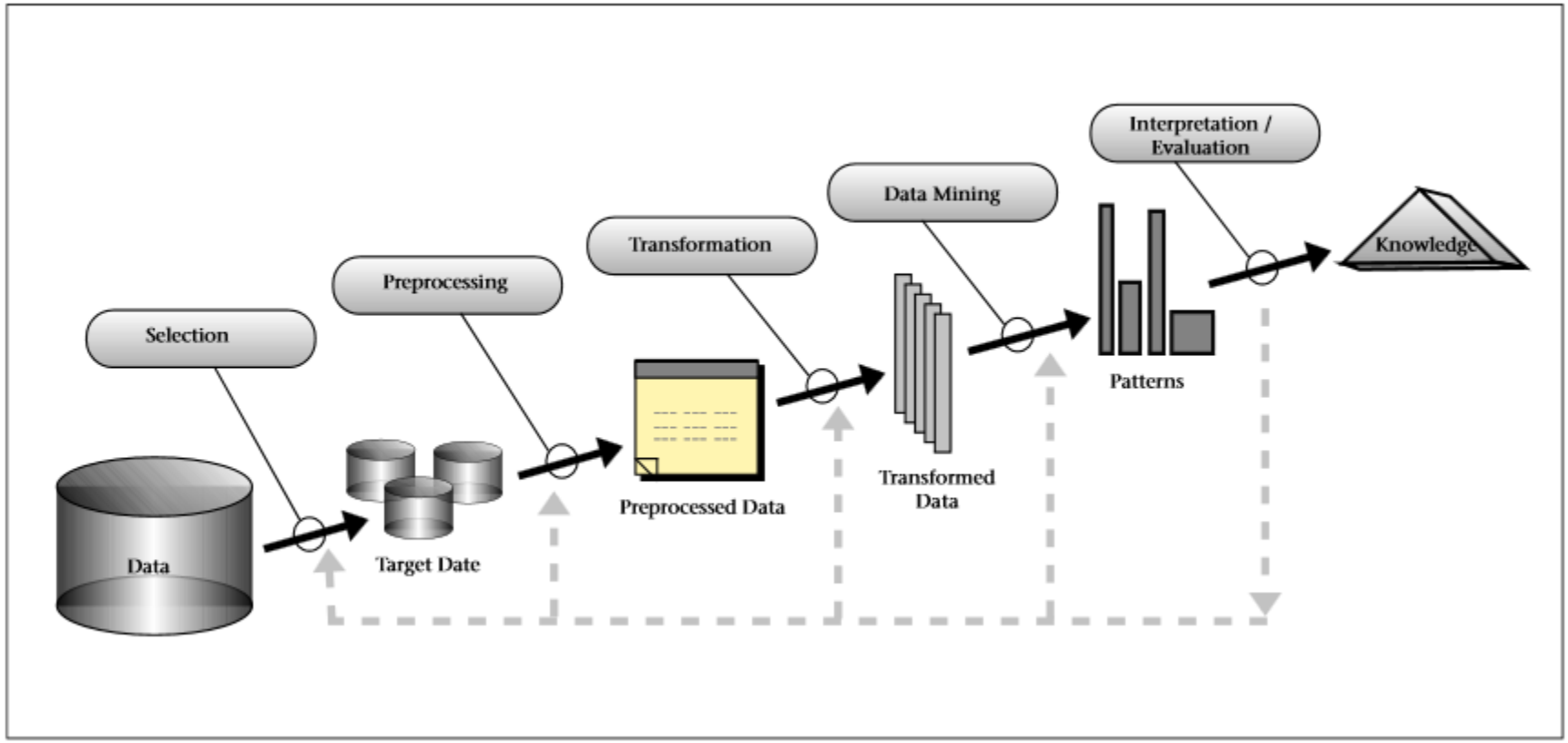


[Sch14]

Gliederung

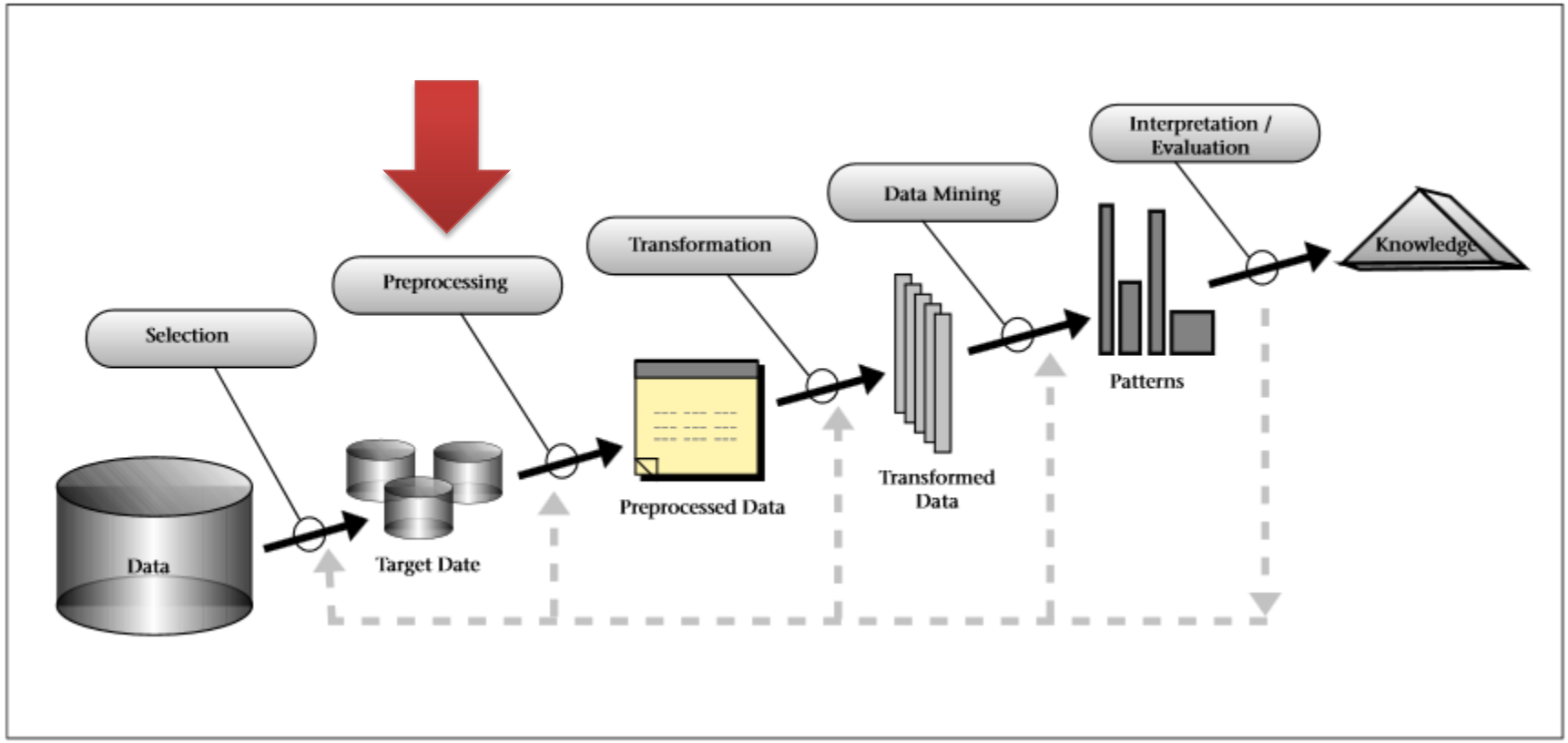
- Motivation
- Textmining
- Tools
- Aktueller Stand
- Ausblick
- Konferenzen & Forschung

Textmining



[FPSS96]

Textmining



[FPSS96]

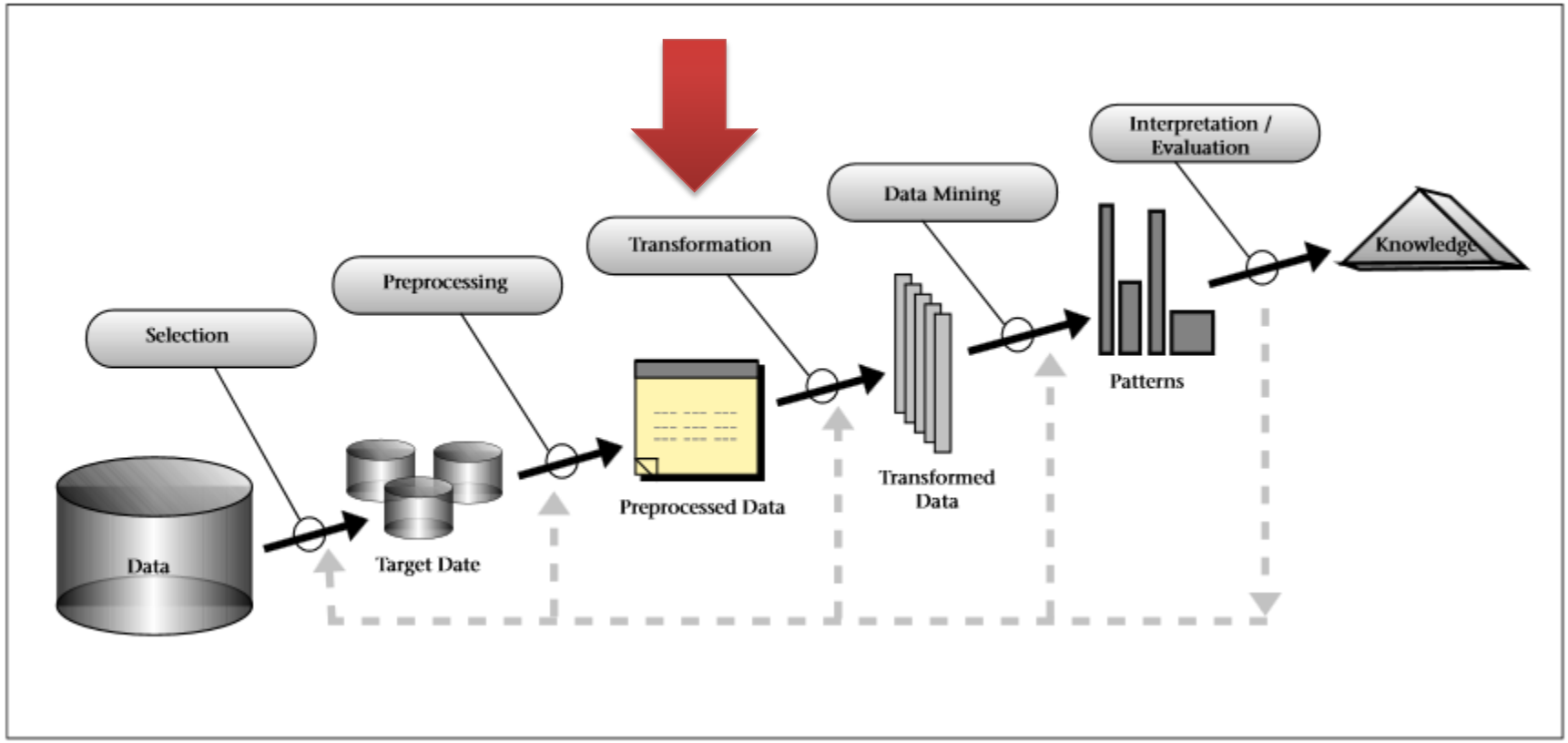
Preprocessing von Dokumenten

- Wie sehen Dokumente aus?
  - Semistrukturierte Text-Dokumente

Preprocessing von Dokumenten

- Normalisierung
- Stemming
- Stopword-Eliminierung

Textmining



[FPSS96]

Transformation

- Dokument „Bag of Words“
- Feature-Vektor über Vector Space Model (VSP)
- Jeder Term als eine Dimension:

	(Apple,	banana,	cat,	window)
doc1 = (	5,	3,	0,	4)
doc2 = (	4,	6,	0,	3)
doc3 = (	0,	3,	7,	5)
doc4 = (	8,	0,	9,	0)
doc5 = (	5,	0,	0,	3)

[LIU12]

- Key-Words Extraction

Distanzfunktion

- Einfache Distanz nach Euklid:

$$dist_{Euklid}(v, w) = \sqrt{\sum_i (v_i - w_i)^2}$$

[CLEV14]

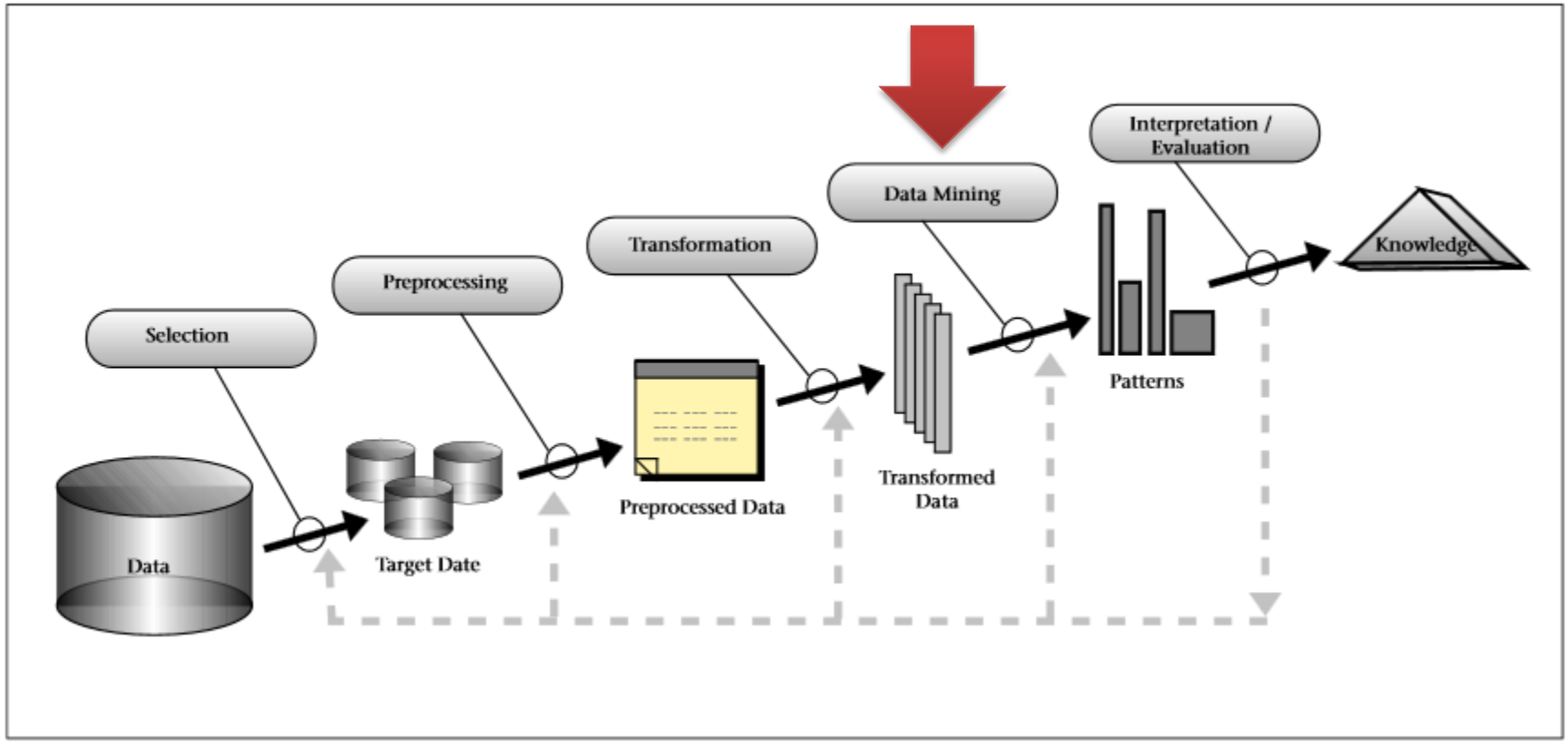
Distanzfunktion (2)

- Cosinus-Ähnlichkeitsmaß
  - Am meisten verbreitet beim Clustering [Feld07]

$$cos(x, y) = \frac{\sum_i x_i * y_i}{\sqrt{\sum_i {x_i}^2 * \sum_i {y_i}^2}}$$

[CLEV14]

Textmining



[FPSS96]

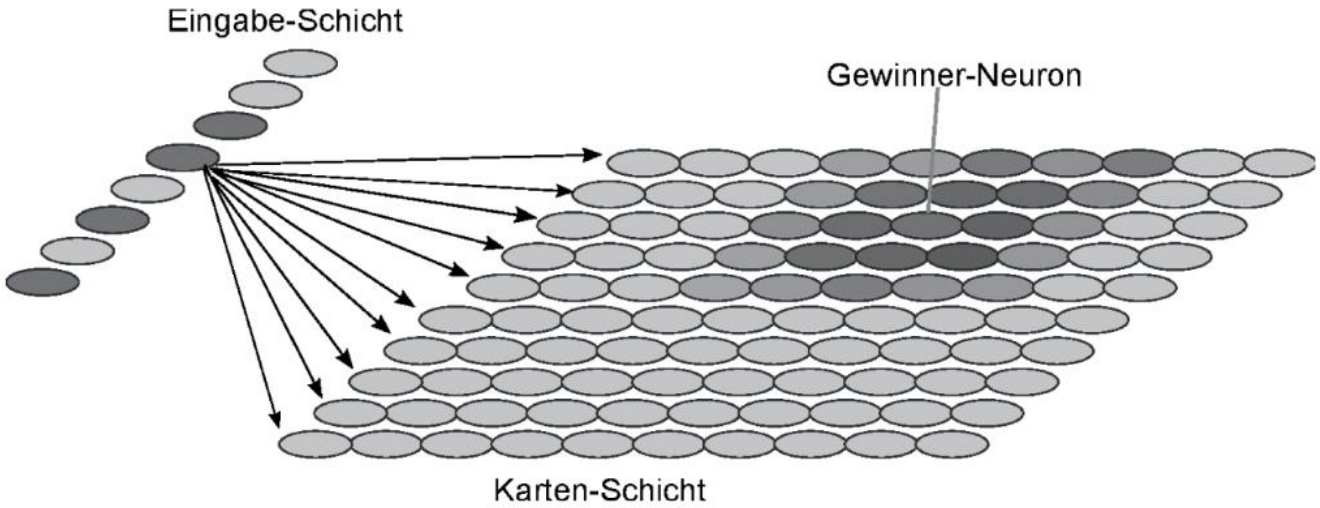
Anwendungsklassen des Data Mining

- Klassifizierung (✓)
- Cluster-Analyse ✓
- Assoziationsanalyse ×
- Numerische Vorhersage ×

Cluster-Analyse

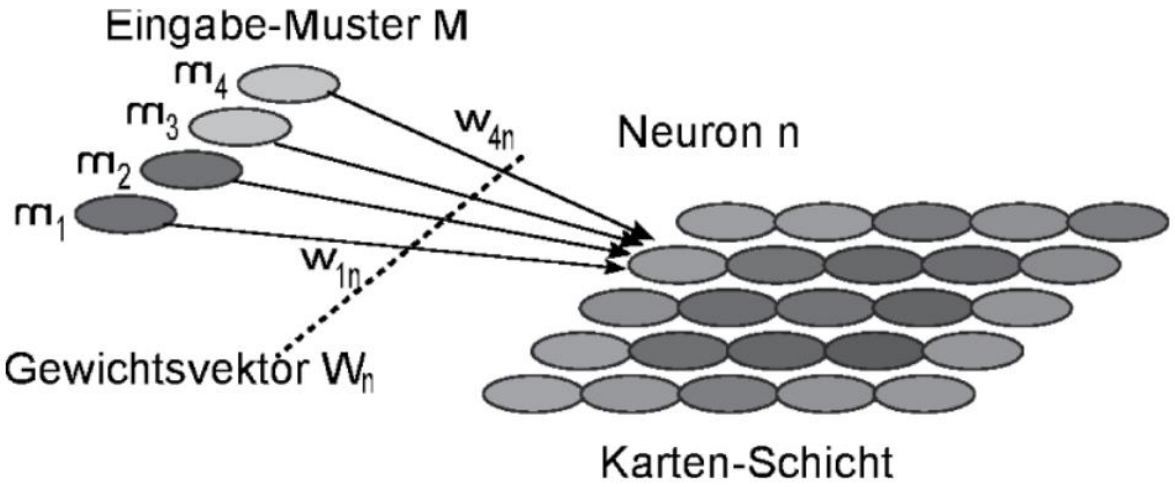
- K-means Algorithmus
  - Einfaches, populäres Verfahren nach MacQueen [Mac67]
- Künstliche neuronale Netze
  - Selbstorganisierte Karten
  - Neuronale Gase
  - ART-Netze

Clusterbildung mittels Self Organizing Map (SOM)



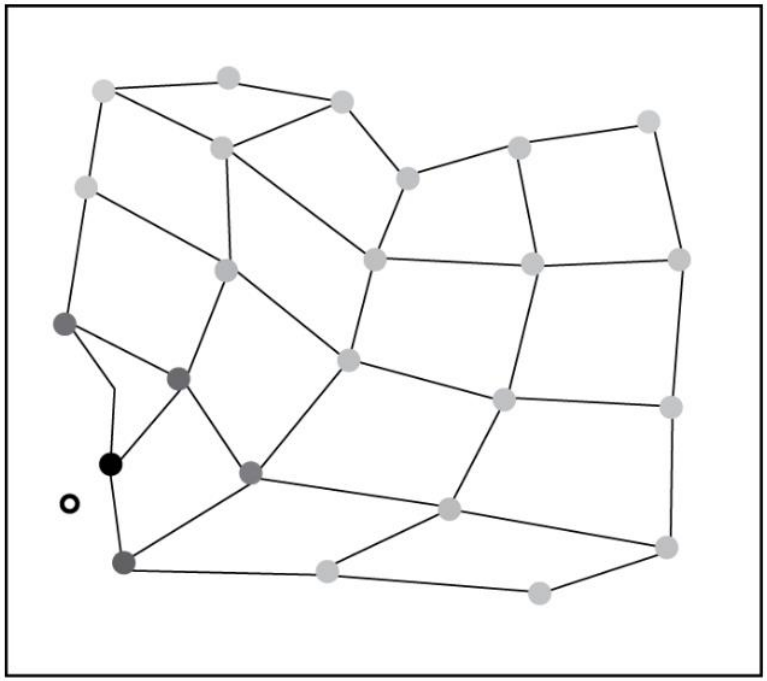
[CLEV14]

SOM (2)


Eingabe-Muster M
m_4
m_3
m_2
m_1
w_{4n}
w_{1n}
Neuron n
Gewichtsvektor W_n
Karten-Schicht

[CLEV14]

SOM (3)



Ring - Eingabe
schwarz - Gewinner-Neuron
grau - Nachbarschaft
hellgrau - unbeeinflusst

[CLEV14]

SOM (4)

Position	Merkmal	Position	Merkmal
1	klein	8	Mähne
2	mittel	9	Federn
3	groß	10	jagt
4	2 Beine	11	rennt
5	4 Beine	12	fliegt
6	Haare	13	schwimmt
7	Hufe		

[CLEV14]

SOM (5)

Tier	1	2	3	4	5	6	7	8	9	10	11	12	13
Taube	1	0	0	1	0	0	0	0	1	0	0	1	0
Henne	1	0	0	1	0	0	0	0	1	0	0	0	0
Ente	1	0	0	1	0	0	0	0	1	0	0	0	1
Gans	1	0	0	1	0	0	0	0	1	0	0	1	1
Eule	1	0	0	1	0	0	0	0	1	1	0	1	0
Falke	1	0	0	1	0	0	0	0	1	1	0	1	0
Adler	0	1	0	1	0	0	0	0	1	1	0	1	0
Fuchs	0	1	0	0	1	1	0	0	0	1	0	0	0
Hund	0	1	0	0	1	1	0	0	0	0	1	0	0
Wolf	0	1	0	0	1	1	0	1	0	1	1	0	0
Katze	1	0	0	0	1	1	0	0	0	1	0	0	0
Tiger	0	0	1	0	1	1	0	0	0	1	1	0	0
Loewe	0	0	1	0	1	1	0	1	0	1	1	0	0
Pferd	0	0	1	0	1	1	1	1	0	0	1	0	0
Zebra	0	0	1	0	1	1	1	1	0	0	1	0	0
Kuh	0	0	1	0	1	1	1	0	0	0	0	0	0

[CLEV14]

SOM (6)

[CLEV14], Tool: SoKo-Wismar (Self-Organizing Kohonen Map)

Gliederung

- Motivation
- Textmining
- Tools
- Aktueller Stand
- Ausblick
- Konferenzen & Forschung

Tools zum Textmining

- Rapidminer (YALE)
- Weka (Waikato Environment for Knowledge Analysis)
- Beagle Search (Apache Lucene)
- Hadoop mit z.B. Apache Tez oder Apache Mahout auf Spark

Gliederung

- Motivation
- Textmining
- Tools
- Aktueller Stand
- Ausblick
- Konferenzen & Forschung

Aktueller Stand

- Vorarbeit durch Marcel Schöneberg (M.-Inf.), Nina Hälker (M.-Next Media)
- Datenbasis Eurozine Netzwerk (www.eurozine.com) [Sch2014]
  - 2700 Journalistische Artikel
  - Meta-Informationen
  - semi-strukturiert in XML (Autor, Abstract, Überschriften usw.)
  - Größtenteils englisch

Aktueller Stand (2)

- Einfache Distanzfunktion
  - Distanz nach Euklid gemäß Gewichtung
- Dossier-Vorschläge anhand ähnlicher Dokumente
- Keine multimedialen Dossiers

Gliederung

- Motivation
- Textmining
- Tools
- Aktueller Stand
- Ausblick
- Konferenzen & Forschung

Ausblick

- Verbesserung der Gewichtung durch Kenntnis der Fachdomäne
- Linguistische Verbesserungen
- Optimierung der Distanzfunktion
- Evaluierung von Methoden zum Clustering
- Entwicklung einer Toolchain

Gliederung

- Motivation
- Textmining
- Tools
- Aktueller Stand
- Ausblick
- Konferenzen & Forschung

Konferenzen

- ACM SIGKKD
  - Knowledge Discovery & Data Mining
- ACM SIGMOD
  - Management of Data
- IEEE Big Data 2014
- ISC Big Data

Forschung (Digital Journalism)

- Center for Digital Journalism
  - Jay Rosen
- Digital Storytelling
  - Bryan Alexander

Quellen

[Clev14] CLEVE, Jürgen; LÄMMEL, Uwe: *Data Mining*. De Gruyter, 2014

[Grim13] GRIMMER, Justin; STEWARD, Brandon M.: *Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts*. Stanford University, 2013

[FPSS96] FAYYAD, Usma M.; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic: *From Data Mining to Knowlege Discovery: An Overview*. In: FYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic; UTHURU-SAMY, Ramasamy (Hrsg.): Advances in Knowlege Discovery and Data Mining. Menlo Park, Cambridge, London: MIT Press, 1996, S. 1-34

[Sch14] SCHÖNEBERG, Marcel: *Automatisierte Erstellung von Pressedossiers durch Textmining*: Kontextualierung im journalistischen Umfeld. 2014 – Masterseminar Ausarbeitung

[Feld07] FELDMAN, Ronen; SANGER, James: *The Text Mining Handbook: Advanced Approaches in Analysing Unstructured Data*. Cambridge University Press, 2007

[Mac67] MACQUEEN, J.: *Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1: 281–97. London, UK: Cambridge University Press 1967.

[Lui12] LIU, Yuan-Chao; LIU, Ming; WANG, Ming: *Application of Self-Organizing Maps in Text Clustering: A Review, Applications of Self-Organizing Maps.* 2012

Vielen Dank für die Aufmerksamkeit!

Gibt es Fragen?