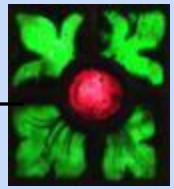
Lehrstuhl für Angewandte Informatik in den
Kultur-, Geschichts- und Geowissenschaften

Otto-Friedrich-Universität Bamberg



Reading Club - Similarity

Similarity Models – An Overview

Sebastian Matyas

09. Juli 2008

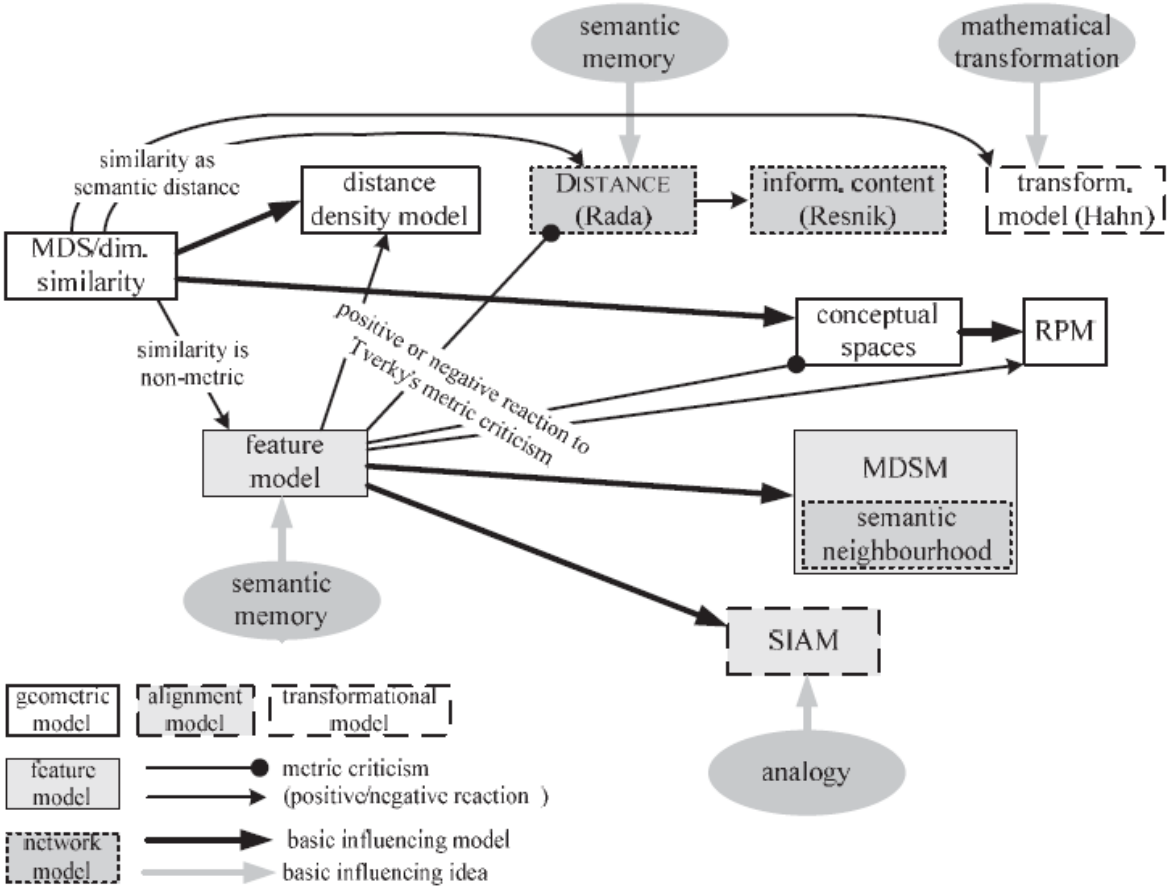


Figure 10 The development of semantic similarity measures

Angela Schwering (2008). Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey, Transactions in GIS *Vol. 12 Issue 1, February 2008*

Geometrisches Modell

- Koordinatensystem
  - Objekte (Instanzen) als Punkte in einem n-dimensionalen Raum (Koordinatensystem)
  - Ähnlichkeit definiert als nicht-negative Zahl einer metrischen Distanzfunktion $\delta(a,b)$
- Axiome
  - Minimality
    $\delta(a,b) \geq \delta(a,a) = 0$
  - Symmetry
    $\delta(a,b) = \delta(b,a)$
  - Triangle inequality
    $\delta(a,b) + \delta(b,c) \geq \delta(a,c)$

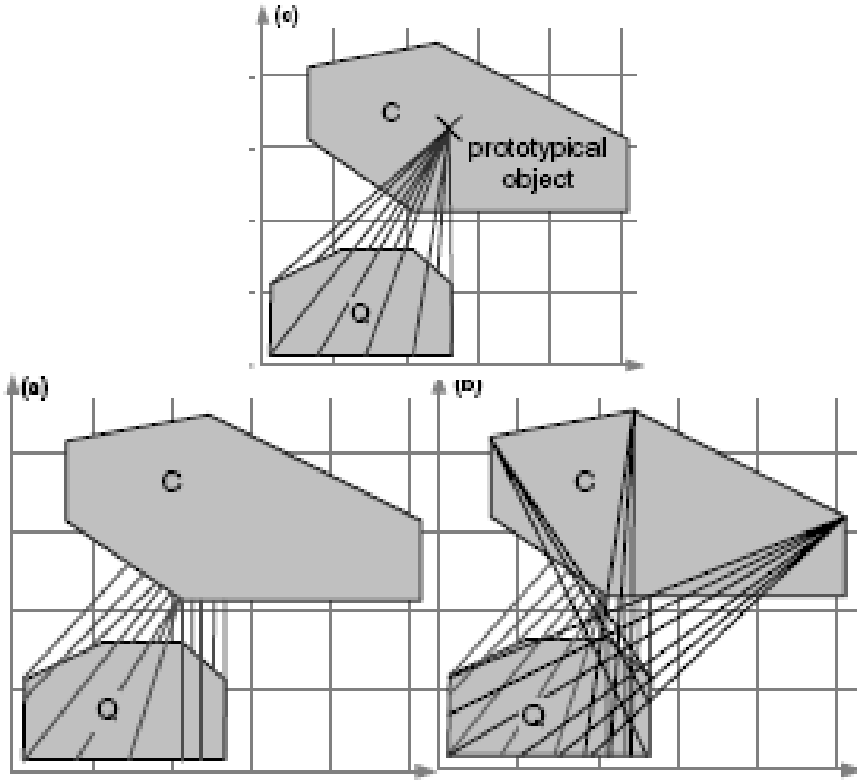
Objekte und Konzepte

- Similarity Models
  - Kann Ähnlichkeit auch zwischen Konzepten berechnet werden?

„ a concept is an idea that characterizes a set or category of objects“ Sloman et al. (1998, p. 192)

Geometrisches Model

- Conceptual Spaces
  - Entwickelt von Gardenförs (2000)
  - Konzepte werden als n-dimensionale konvexe Bereiche repräsentiert
  - Ähnlichkeit zwischen Konzepten z.B. anhand von prototypischen Instanzen (siehe Prototype Theory – Rosch 1975)



Ähnlichkeitsbestimmung zwischen Konzepten aus Schwering und Rubal (2005)

A. Schwering and M. Raubal (2005), Measuring Semantic Similarity Between Geospatial Conceptual Regions, LNCS 3799, GeoSpatial Semantics 2005, pp.90-106

Teil 1
Feature Model

Teil 2
Network Model

Teil 3
Alignment – Transformational Model

Kritik

- Axiome
  - Minimality
    $\delta(a,b) \geq \delta(a,a) = 0$
  - Symmetry
    $\delta(a,b) = \delta(b,a)$
  - Triangle inequality
    $\delta(a,b) + \delta(b,c) \geq \delta(a,c)$
- Gegenbeispiele
  - Gilmore, Hersh, Camarazza and Griffin (1979) Buchstabe M wurde öfter als H erkannt als als M
  - Tversky, 1977: „North Korea is like Red China“ - „Red China is like North Korea
  - Tversky , 1977: Jamaica is similar to Cuba (*geographisch)*; Cuba is similar to Russia (*politisch*), but Jamaica and Russia are not similar at all

Keine geometrische Dimensionen…

?
=

Feature Matching

- Grundannahmen
  - **Matching**:

$$s(a,b) = F(A \cap B, A - B, B - A)$$

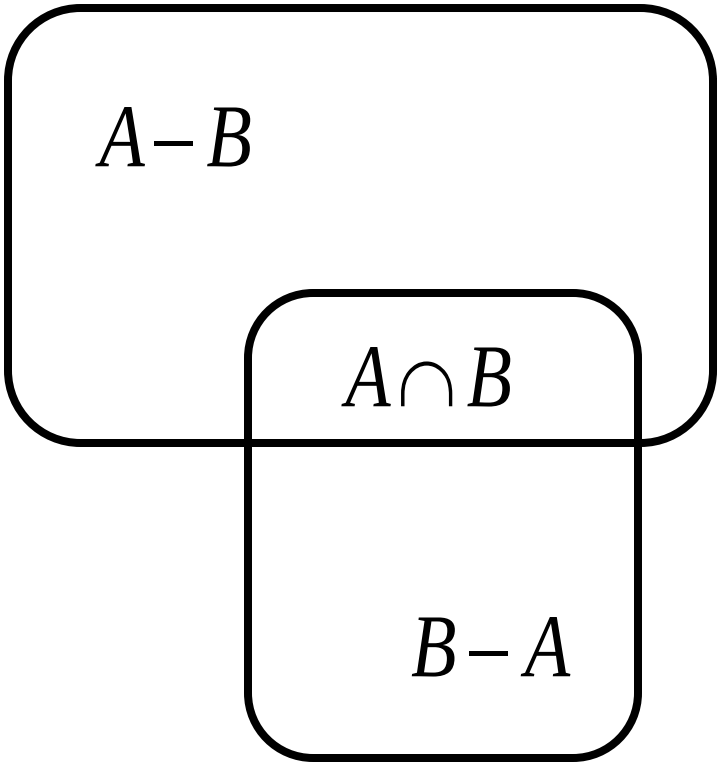
  - **Monotonie**:

$$s(a,b) \geq s(a,c)$$

$$A \cap B \supset A \cap C, A - B \subset A - C$$

$$B - A \subset C - A$$

  - Sowie **Unabhängigkeit**, **Invarianz** und **Lösbarkeit**



Contrast und Ratio Model

- Contrast Model

$$S(a,b) = \theta * f(A \cap B) + \alpha * f(A - B) + \beta * f(B - A)$$

- Ratio Model

$$S(a,b) = \frac{f(A \cap B)}{f(A \cap B) + \alpha * f(A - B) + \beta * f(B - A)}$$

$$\alpha, \beta, \theta \geq 0$$

Teil 1
Feature Model

Teil 2
Network Model

Teil 3
Alignment – Transformational Model

Semantische Netze

- Wissensrepräsentation
  - Graph-basierte Darstellung
  - Kanten stellen Beziehungen zwischen Knoten dar (Konzepte, bzw. Objekte)
  - Taxonomische, hyponomische und partonomische oder andere hierarchische Beziehungen
- Grundannahmen
  - **Lösbarkeit**: Verbindung zwischen Konzepte/Objekte muss vorhanden sein
  - **Vergleichbarkeit:** Verbindungen sind alle gleichgewichtet

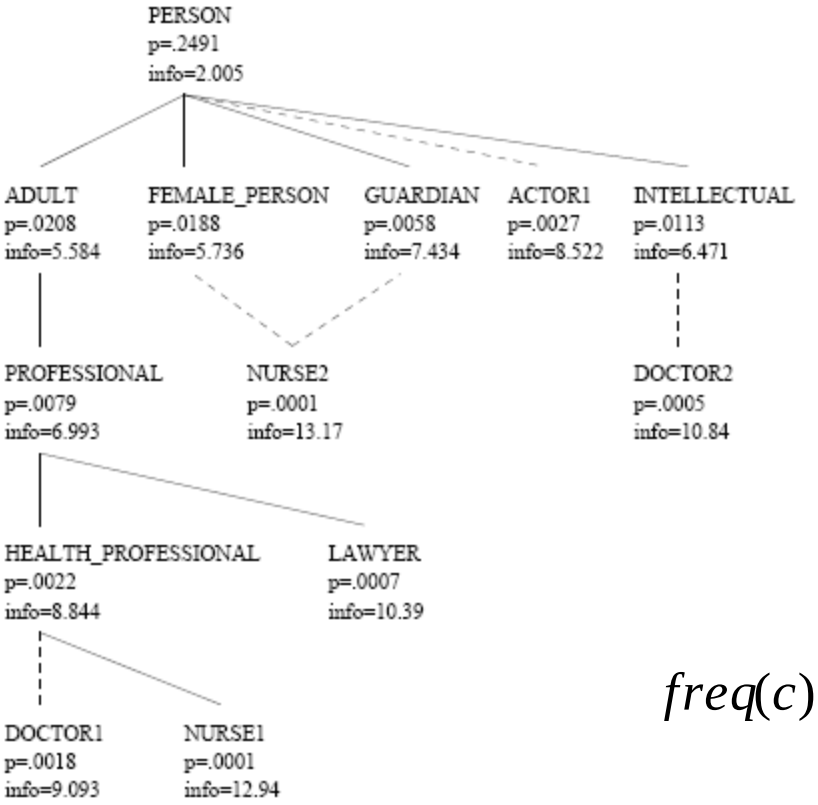
    (DISTANCE Maß von Rada et el. 1989: „*average minimum path length over all par-wise combinations of nodes between two subsets of nodes*“)

Information Content

- Ähnlichkeitsmaß
  - Entfernung im Semantischen Netz
  - Verbindungen sind nicht gleich gewichtet
  - Analog zur Informationstheorie (Ross 1998), definiert Resnik (1995) den *Information Content* eines Konzeptes als den negativen Logarithmus seiner Wahrscheinlichkeit
  - Die Ähnlichkeit von zwei Konzepten c1 und c2 ist der maximale *Information Content* von allen Konzepten die c1 und c2 subsumieren

$$sim(c_1, c_2) = \max_{c \in S(c_1, c_2)} \left[-\log\ p(c) \right]$$

Beispiel



- WordNet Taxonomy
  - Häufigkeiten von Konzepten repräsentiert als Nomen in einem Corpus
  - Jedes Nomen wird als Vorkommen der taxonomischen Klasse gezählt, welches es enthält

$$freq(c) = \sum_{n \in words(c)} count(n) \qquad p(c) = \frac{freq(c)}{N}$$

Philip Resnik (1999). Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language, *Journal of Artificial Intelligence Research 11*, pp. 95-130

Matching-Distance Similarity Measure (MDSM)

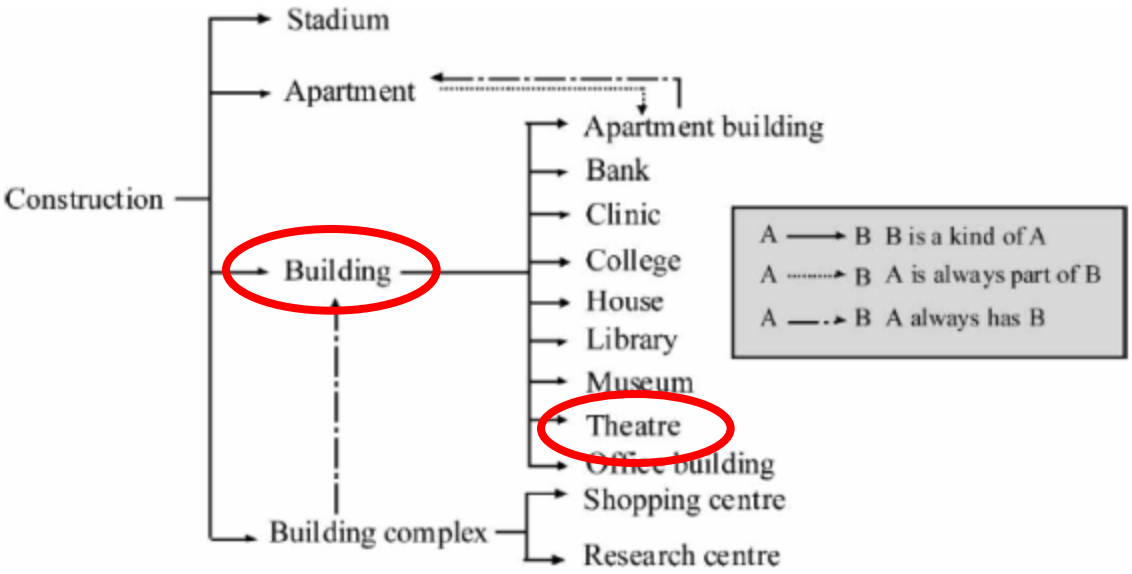
- Network und Feature Model

$$S(c_1, c_2) = \omega_p \cdot S_p(c_1, c_2) + \omega_f \cdot S_f(c_1, c_2) + \omega_a \cdot S_a(c_1, c_2)$$

$$S_t(c_1, c_2) = \frac{|C_1 \cap C_2|}{|C_1 \cap C_2| + \alpha(c_1, c_2) \cdot |C_1 - C_2| + (1 - \alpha(c_1, c_2)) \cdot |C_2 - C_1|}$$

  - t … type of feature (part, attribute, function)
  - c1, c2 … compared entity classes
  - C1, C2 … respective sets of features of type t for c1, c2

$$\alpha(c_1, c_2) = \begin{cases} \dfrac{d(c_1, \text{lub})}{d(c_1, \text{lub}) + d(c_2, \text{lub})}, d(c_1, \text{lub}) \leq d(c_2, \text{lub}) \\ 1 - \dfrac{d(c_1, \text{lub})}{d(c_1, \text{lub}) + d(c_2, \text{lub})}, d(c_1, \text{lub}) > d(c_2, \text{lub}) \end{cases}$$



M. Andrea Rodriguez; Max J. Egenhofer, (2004) Comparing geospatial entity classes: an asymmetric and context-dependent similarity measure, International Journal of Geographical Information Science, Volume 18, Issue 3, pages 229 - 256

Beispiel (1)

Entity class	Parts	Functions	Attributes
Building	Foundation		Architectural properties
	Roof		External material construction
	Wall		Height
			Location
			Name
			Owner type
			Structure type
			User type
Theatre	Dressing room	Perform	Architectural properties
	Entrance hall	Present	External material construction
	Foundation	Recreate	Height
	Orchestra		Location
	Roof		Name
	Spectator stands		Owner type
	Stage		Structure type
	Ticket office		User type
	Wall		

Beispiel (2)

- Theatre – building
  - depth (theatre) [1] > depth (building) [0]
    => α = 1 – 1 / (1+0) = 0
  - S_p = 3 / (3 + 0 + 0) = 1
  - S_f = 0 (no functions for building)
  - S_a = 1 (same attributes)
- Building – theatre
  - depth (building) [0] < depth (theatre) [1]
    => α = 0 / (1+0) = 0
  - S_p = 3 / (3 + 0 + 6) = 1/3
  - S_f = 0 (no functions for building)
  - S_a = 1 (same attributes)

Beispiel (3)

Entity classes	α	S_p	S_f	S_a	$S(c_1, c_2)$
theatre, building	0.0	1.0	0.0	1.0	**0.66**
building, theatre	0.0	0.33	0.0	1.0	**0.44**

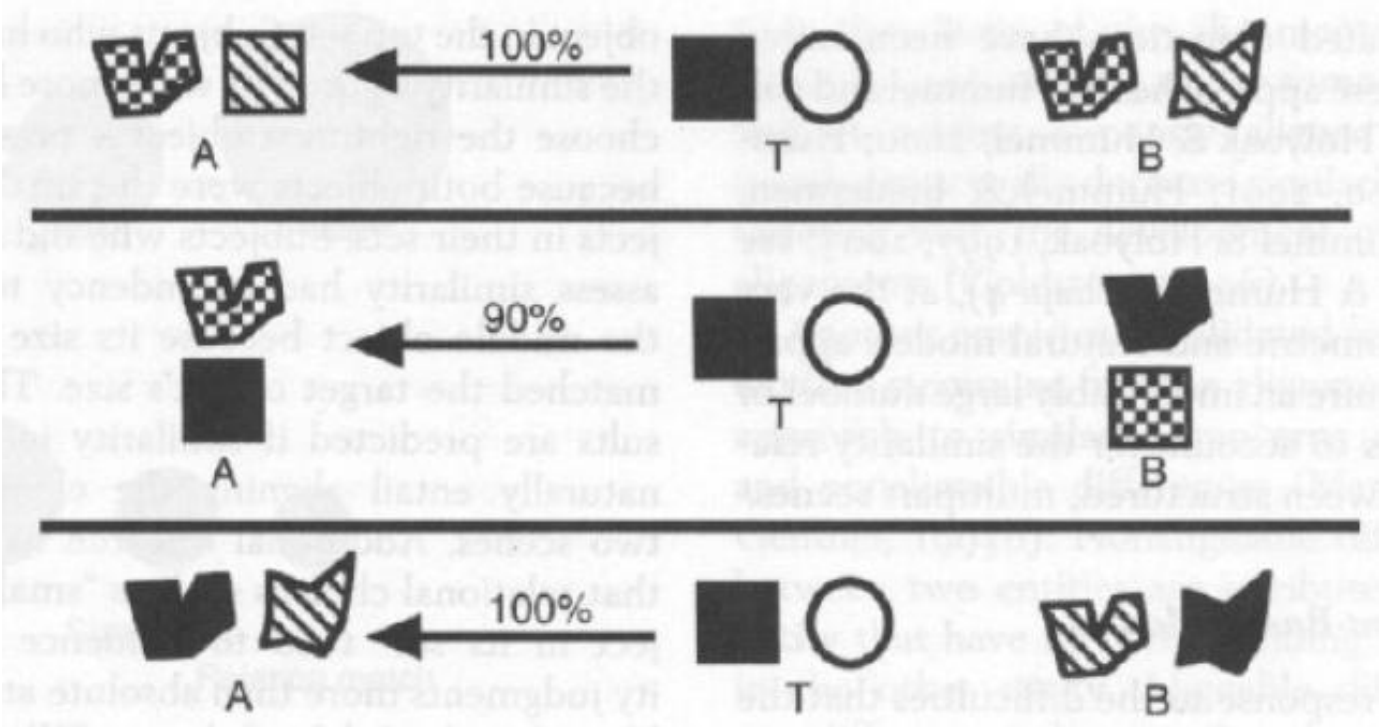
$$\omega_p, \omega_f, \omega_a = 1/3$$

Teil 1
Feature Model

Teil 2
Network Model

Teil 3
Alignment – Transformational Model

Nur Features?

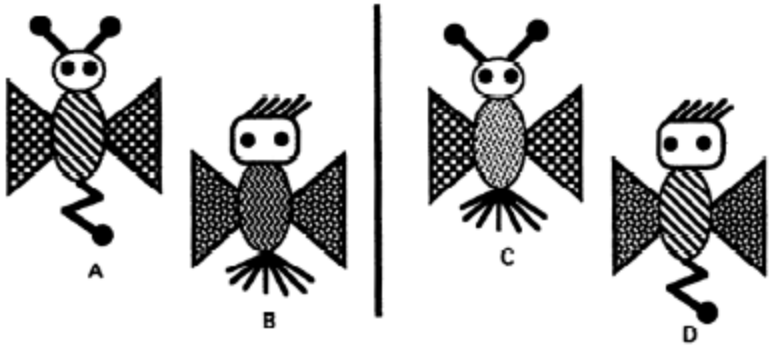


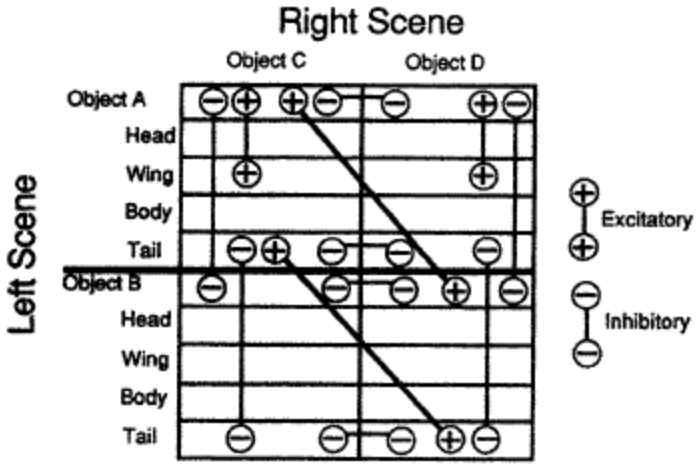
Goldstone, R. L, & Son, J. (2005). Similarity. In K. Holyoak & R. Morrison (Eds.). *Cambridge Handbook of Thinking and Reasoning*. Cambridge: Cambridge University Press. (pp. 13-36)

Alignment Model

- Grundannahmen
  - **Homogenität**: Homogene Struktur der zu vergleichenden Objekte
  - **Vergleichbarkeit**: Jedes Element hat denselben Einfluss auf das Ähnlichkeitsmaß
  - Sowie **Lösbarkeit** und **Unabhängigkeit**
- SIAM
  - Similarity as Interactive Activation and Mapping
  - Inspiriert durch Arbeiten zu analogical reasoning (Genter, 1983) und interactive activation models of perception (McClelland & Elman, 1986)
  - Nodes (=Hypothesen) sind zentrale Elemente
  - Feature-to-Feature, Object-to-Object und Role-to-Role nodes

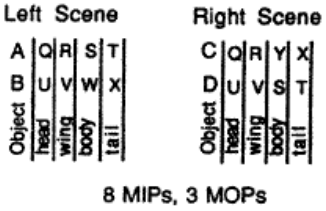
Beispiel





Right Scene

Left Scene	Object C	Object D
Object A	0.5/0.5/0.501/0.561	0.5/0.5/0.499/0.439
Head	(1.0)/0.55/0.609/0.673/0.976	(0.0)/0.45/0.392/0.327/0.024
Wing	(1.0)/0.55/0.609/0.673/0.976	(0.0)/0.45/0.392/0.327/0.024
Body	(0.0)/0.45/0.402/0.355/0.097	(1.0)/0.55/0.599/0.645/0.903
Tail	(0.0)/0.45/0.392/0.327/0.045	(1.0)/0.55/0.609/.673/0.955
Object B	0.5/0.493/0.480/0.353	0.5/0.5/0.501/0.561
Head	(0.0)/0.45/0.392/0.326/0.021	(1.0)/0.55/0.609/0.673/0.976
Wing	(0.0)/0.45/0.392/0.326/0.021	(1.0)/0.55/0.609/0.673/0.976
Body	(0.0)/0.45/0.422/0.413/0.595	(0.0)/0.45/0.402/0.355/0.097
Tail	(1.0)/0.55/0.607/0.672/0.922	(0.0)/0.45/0.392/0.327/0.045



8 MIPs, 3 MOPs

MIP = Match in Place
MOP = Match out of Place

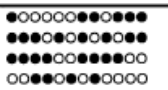
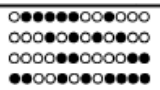
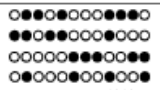
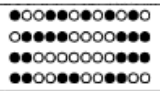
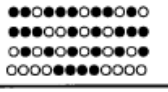
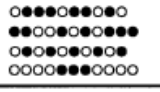
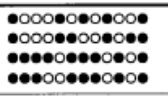
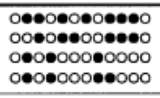
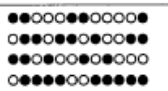
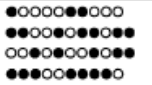
Goldstone, R. L. (1994). Similarity, Interactive Activation, and Mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 3- 28

Transformational Model

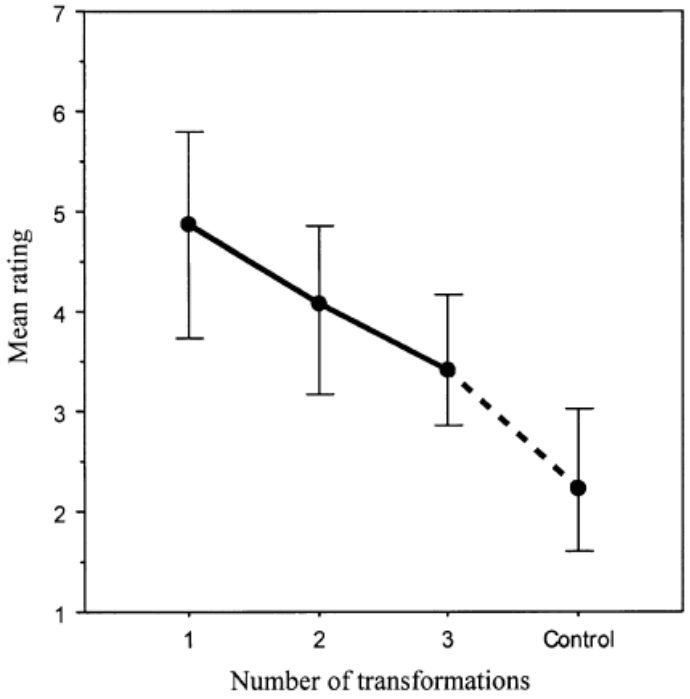
- Grundannahmen
  - **Vergleichbarkeit**: Jede Transformation darf die Ähnlichkeit zwischen Objekten nur um dasselbe Maß beeinflussen
  - **Komplexität**: Einfache Konzepte erfordern nur einfache Transformation aber komplexe Konzepte auch komplexe Transformationen
  - Sowie **Lösbarkeit**
- Ähnlichkeitsmaß
  - Basiert auf dem Ansatz des Representational Distortion (RD) – (Chater & Huhn, 1997)
  - Kolmagorov complexity theory (Li & Vitáyi, 1997)
  - *Intuitiv*: Komplexität einer Representation, x, ist die Länge des kürzesten Computerprogramms, welches die Repräsentation generieren kann

Experimenteller Beweis

Table 1
Stimuli used in Experiment 1

No. of transformations	Type	Stimuli	
		Item one	Item two
1	Reversal		
1	Mirror		
1	Phasic		
1	Deletion		
2	Reversal & Mirror		
2	Deletion & Mirror		

...



Hahn, U., Richardson, L.B. & Chater, N., (2001). Similarity: A transformational approach. In Proceedings of the 23rd Annual Conference of the Cognitive Science Society, pp. 393-398

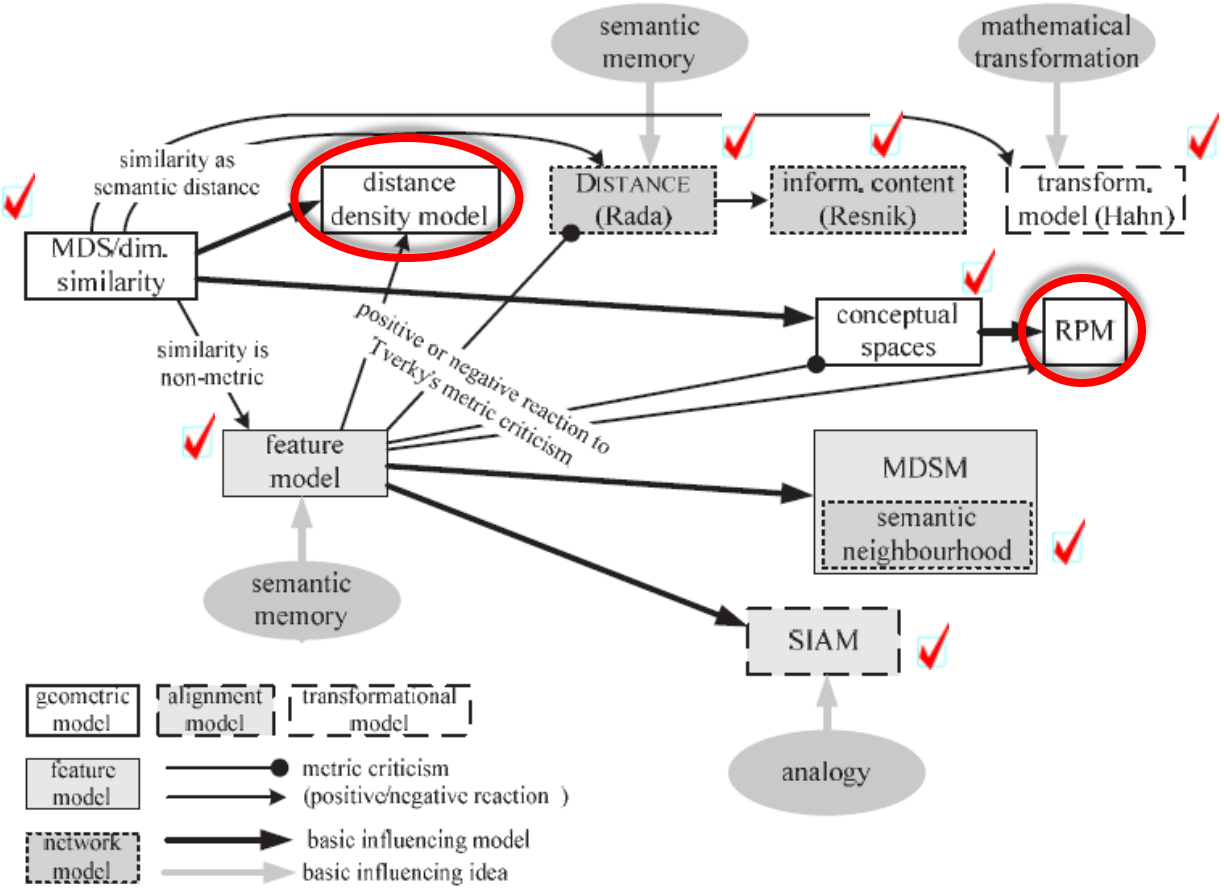
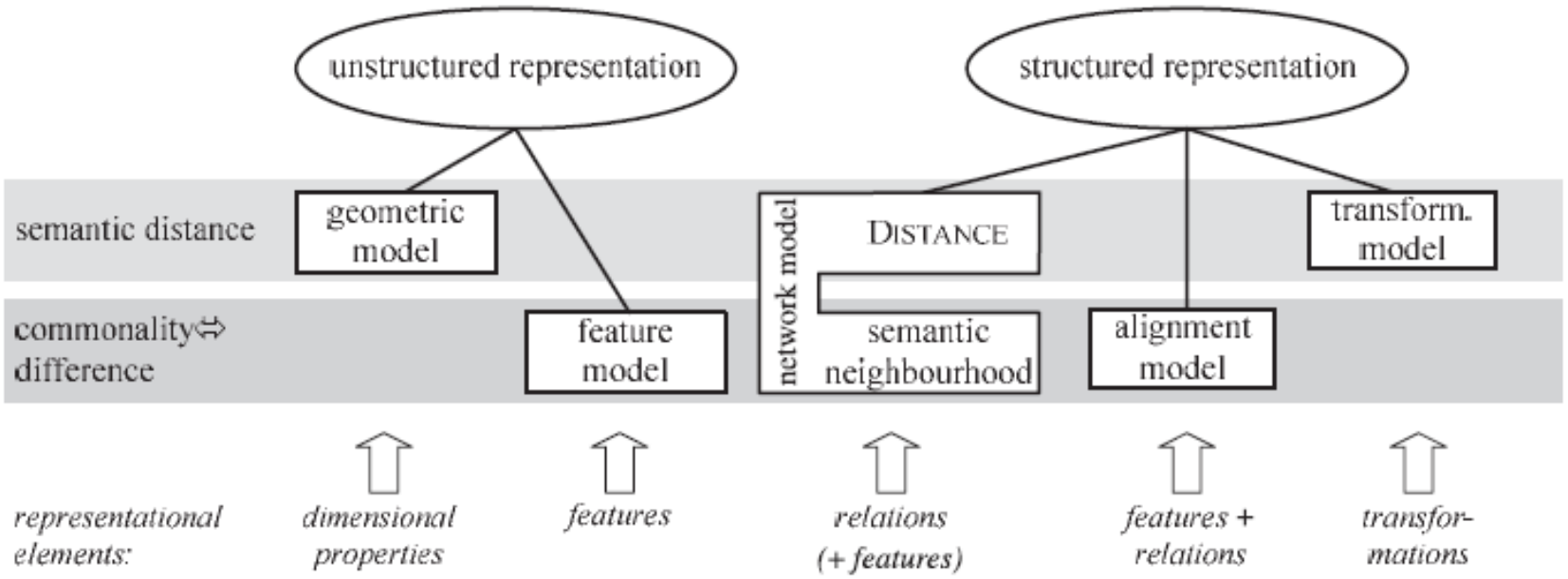


Figure 10 The development of semantic similarity measures

Angela Schwering (2008). Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey, Transactions in GIS *Vol. 12 Issue 1, February 2008*

Universales Ähnlichkeitsmodell?



Angela Schwering (2008). Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey, Transactions in GIS *Vol. 12 Issue 1, February 2008*

Eins noch…

http://www.similarity-blog.de/