



## Ähnlichkeitsmaße für Vektoren

Karin Haenelt

25.10.2012

# Inhalt

- Einführung
  - Ähnlichkeitsmaß
  - Ähnlichkeitsbetrachtungen
- Gebräuchliche Ähnlichkeitsmaße im Information Retrieval
  - Korrelationsmaße: Einfache Methode, Cosinus, Dice-Koeffizient, Jaccard-Koeffizient, Overlap-Koeffizient
  - Distanzmaße: Euklidische Distanz
- Eine Analyse der Ähnlichkeitsmaße von Jones/Furnas (1987)
- Beispiel 1: Berechnung der Ähnlichkeitsmaße für sechs Dokumentvektoren
- Beispiel 2: Bestimmung der Ähnlichkeit von Nomina auf der Basis von Prädikat-Objekt-Kookkurrenz-Paaren

# Ähnlichkeitsmaße für Vektoren

## Bestimmung

- geben für jeweils zwei Vektoren einen numerischen Wert, der die Ähnlichkeit zwischen den Vektoren angibt
- verschiedene Maße versuchen verschiedene Aspekte der Ähnlichkeit zu manifestieren

# Ähnlichkeitsbetrachtungen

## Verhältnisse von Termgewichten

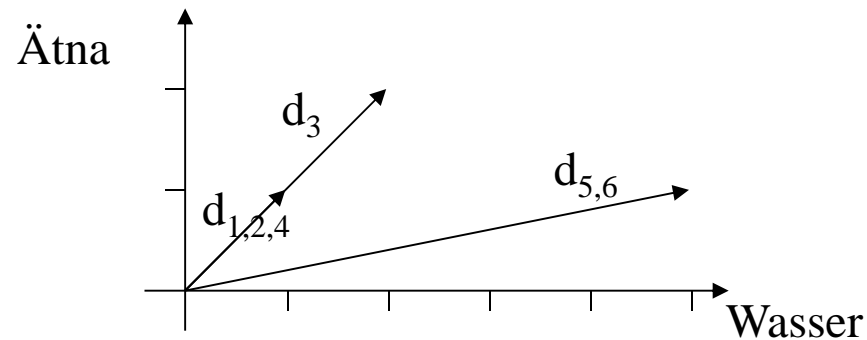
	d 1	d 2	d 3	d 4	d 5	d 6
Ätna	1	1	2	1	1	1
Vesuv	1	1	2	0	2	0
Stromboli	1	1	2	1	3	3
Feuer	1	1	2	0	4	0
Wasser	1	1	2	1	5	5
Lava	1	1	2	0	6	0

- objektintern
  - Verhältnis von  $\text{Term}_i$  zu den anderen Termen eines Dokuments
  - **Wichtigkeit eines Terms** für ein Objekt
  - Hinweise auf **semantischen Inhalt** oder **Themengebiet**
- objektübergreifend
  - Relevanz von  $\text{Dokument}_j$  für  $\text{Term}_i$

# Ähnlichkeitsbetrachtungen

## Interpretation von Term-Vektoren im Information Retrieval

	d 1	d 2	d 3	d 4	d 5	d 6
Ätna	1	1	2	1	1	1
Wasser	1	1	2	1	5	5



- **Richtung**
  - bestimmt durch objektinternes Verhältnis der Terme
  - möglicherweise Hinweis auf **Thema**
- **Länge** (im Verhältnis zu anderen Vektoren)
  - bestimmt durch objektübergreifendes Verhältnis der Termgewichte
  - möglicherweise Hinweis auf **Intensität eines Themas**

Jones/Furnas, 1987

# Inhalt

- Einführung
  - Ähnlichkeitsmaß
  - Ähnlichkeitsbetrachtungen
- **Gebräuchliche Ähnlichkeitsmaße im Information Retrieval**
  - Korrelationsmaße: Einfache Methode, Cosinus, Dice-Koeffizient, Jaccard-Koeffizient, Overlap-Koeffizient
  - Distanzmaße: Euklidische Distanz
- Eine Analyse der Ähnlichkeitsmaße von Jones/Furnas (1987)
- Beispiel 1: Berechnung der Ähnlichkeitsmaße für sechs Dokumentvektoren
- Beispiel 2: Bestimmung der Ähnlichkeit von Nomina auf der Basis von Prädikat-Objekt-Kookkurrenz-Paaren

# Ähnlichkeitsmaße für Vektoren

- **Korrelationsartige Maße:** größter Wert entspricht dem ähnlichsten Paar
  - Cosinus des Winkels zwischen Vektoren
  - Dice-Koeffizient
  - Jaccard-Koeffizient
  - Overlap-Koeffizient
- **Distanz-Maße:** kleinster Wert entspricht dem ähnlichsten Paar
  - Euklidische Distanz

---

(Anderberg, 1973, 134)

# Ähnlichkeitsmaße im IR

	Binäre Vektoren <sup>1)</sup>	Vektoren mit reellen Werten <sup>2)</sup>
Einfache Übereinstimmg.	$ X \cap Y $	$\sum_{k=1}^{\text{\#Dimensionen}} (weight_{xk})(weight_{yk})$
Cosinus-Koeffizient	$\frac{ X \cap Y }{\sqrt{ X  \times  Y }}$	$\frac{\sum_{k=1}^n weight_{xk} \cdot weight_{yk}}{\sqrt{\sum_{k=1}^n weight_{xk}^2} \cdot \sqrt{\sum_{k=1}^n weight_{yk}^2}}$
Dice-Koeffizient	$\frac{2 X \cap Y }{ X  +  Y }$	$\frac{2 \sum_{k=1}^n (weight_{xk} \cdot weight_{yk})}{\sum_{k=1}^n weight_{xk} + \sum_{k=1}^n weight_{yk}}$
Jaccard (oder Tanimoto)-Koeffizient	$\frac{ X \cap Y }{ X \cup Y }$	$\frac{\sum_{k=1}^n (weight_{xk} \cdot weight_{yk})}{\sum_{k=1}^n weight_{xk} + \sum_{k=1}^n weight_{yk} - \sum_{k=1}^n (weight_{xk} \cdot weight_{yk})}$
Overlap-Koeffizient	$\frac{ X \cap Y }{\min( X ,  Y )}$	$\frac{\sum_{k=1}^n \min(weight_{xk}, weight_{yk})}{\min(\sum_{k=1}^n weight_{xk}, \sum_{k=1}^n weight_{yk})}$

|X| steht für die Anzahl der Nicht-Null-Werte im binären Vektor



# Inhalt

- Einführung
  - Ähnlichkeitsmaß
  - Ähnlichkeitsbetrachtungen
- Gebräuchliche Ähnlichkeitsmaße im Information Retrieval
  - Korrelationsmaße: Einfache Methode, Cosinus, Dice-Koeffizient, Jaccard-Koeffizient, Overlap-Koeffizient
  - Distanzmaße: Euklidische Distanz
- Eine Analyse der Ähnlichkeitsmaße von Jones/Furnas (1987)
- Beispiel 1: Berechnung der Ähnlichkeitsmaße für sechs Dokumentvektoren
- Beispiel 2: Bestimmung der Ähnlichkeit von Nomina auf der Basis von Prädikat-Objekt-Kookkurrenz-Paaren

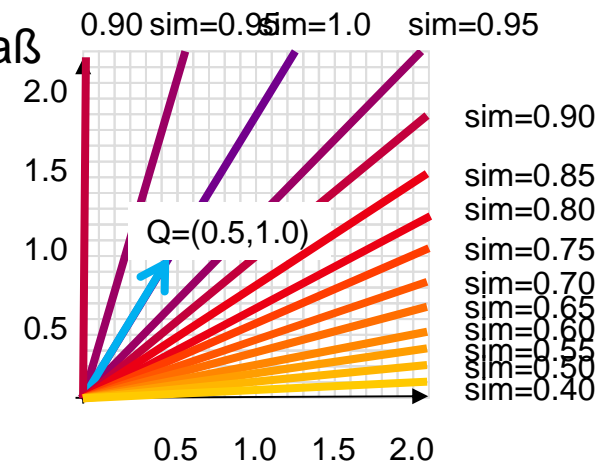
# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

- William P. Jones und George W. Furnas (1987). Pictures of Relevance: A Geometric Analysis of Similarity Measures. In: *Journal of the American Society for Information Science*. 38 (6), S. 420-442.
- Vergleich der Ähnlichkeitsmaße durch geometrische Interpretation der Vektoren und Analyse

# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

## Untersuchungsmethode

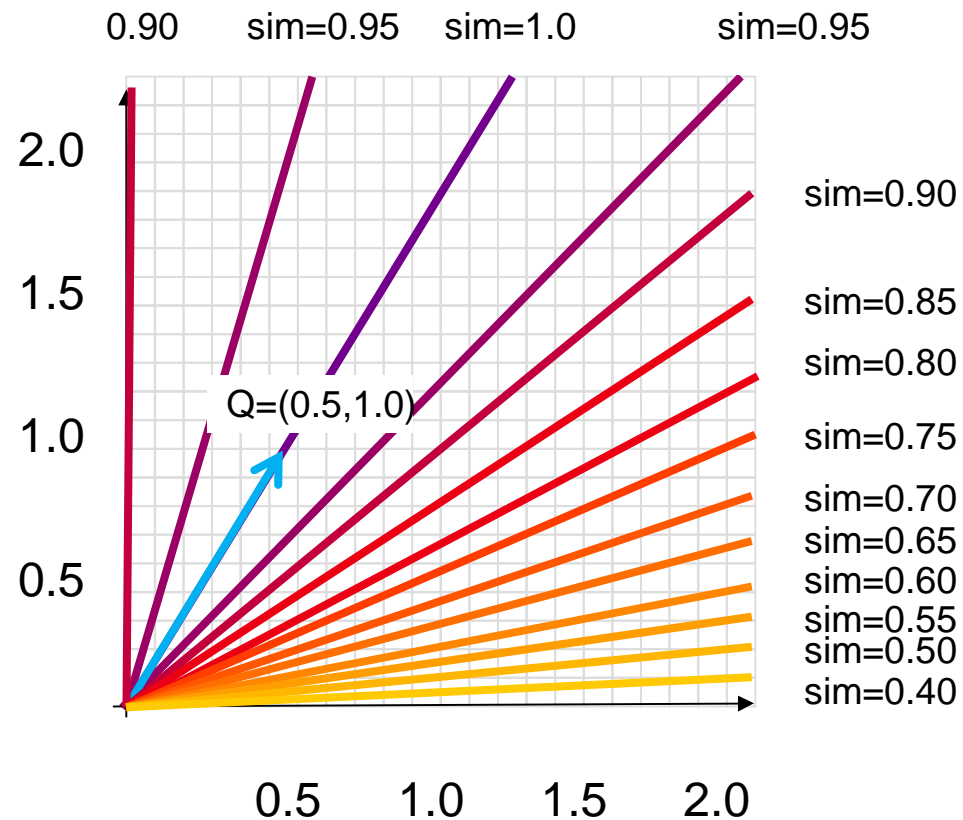
- exemplarische Untersuchung im zweidimensionalen Raum (ermöglicht geometrische Interpretation)
- fester Query-Vektor als Referenzvektor
- Kartierung der Ähnlichkeitswerte zum Referenzvektor
  - Ähnlichkeitswerte: Werte, die ein Maß den anderen Punkten in der Ebene zuweist
  - ein Punkt repräsentiert die Pfeilspitze eines Vektors
- Verbindung der Punkte mit gleichen Ähnlichkeitswerten
  - es ergeben sich Konturlinien: **iso-similarity contours**
  - analog zu Höhenlinien in der Geographie



# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

## Untersuchungsmethode

### Beispiel: Iso-Similarity-Konturen des Cosinus-Maßes



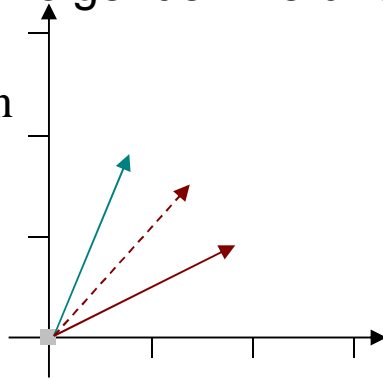
Jones/Furnas, 1987

# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

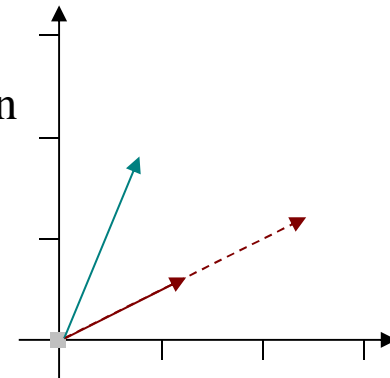
## Untersuchte Eigenschaften der Ähnlichkeitsmaße

- Beschreibung der Veränderung der Werte an Hand der Iso-Similarity-Konturen bei folgenden Veränderungen:

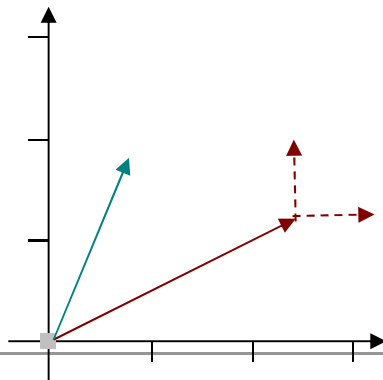
Veränderung des Winkels zwischen den Vektoren



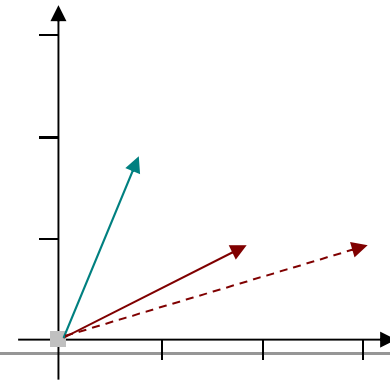
Veränderung des Radius der Vektoren (Länge)



Addition beliebiger Komponenten-Werte



Vergrößerung eines Einzelwertes



# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

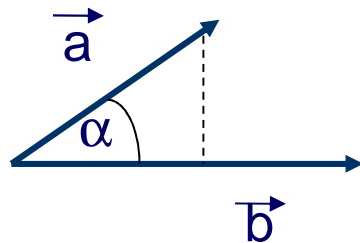
## Untersuchte Eigenschaften der Ähnlichkeitsmaße

- Ähnlichkeitsmaße
    - haben unterschiedliche Isokonturen
    - reflektieren unterschiedliche Vorzüge des Maße bezüglich
      - Richtung („Thema“)
      - Länge („Intensität des Themas“)
  - Ausgangspunkt: algebraische Analyse der Ähnlichkeitsmaße
  - Ziel: „semantische Analyse“ der Ähnlichkeitsmaße
  
  - Untersuchungsfragen
    - bei welcher Veränderung ändern sich die Werte?
    - ändern sich die Werte monoton?
-

# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

## Skalarprodukt

- Skalarprodukt: Multiplikation der Beträge zweier Vektoren unter Berücksichtigung der Richtungsabhängigkeit der Vektoren
- geometrische Darstellung



$$\vec{a} \cdot \vec{b} = |\vec{a}| |\vec{b}| \cos \alpha$$

- algebraische Darstellung

$$\sum_{i=1}^n a_i \times b_i$$

# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

## Skalarprodukt

	Binäre Vektoren	Vektoren mit reellen Werten
Einfache Übereinstimmg.	$ X \cap Y $	$\sum_{k=1}^{\#Dimensionen} (x_k)(y_k)$

- Binäre Vektoren:  
zählt Anzahl der Dimensionen,  
in denen die Werte  
beider Vektoren  $\neq 0$

- Vektoren mit reellen Werten:

Beispiel

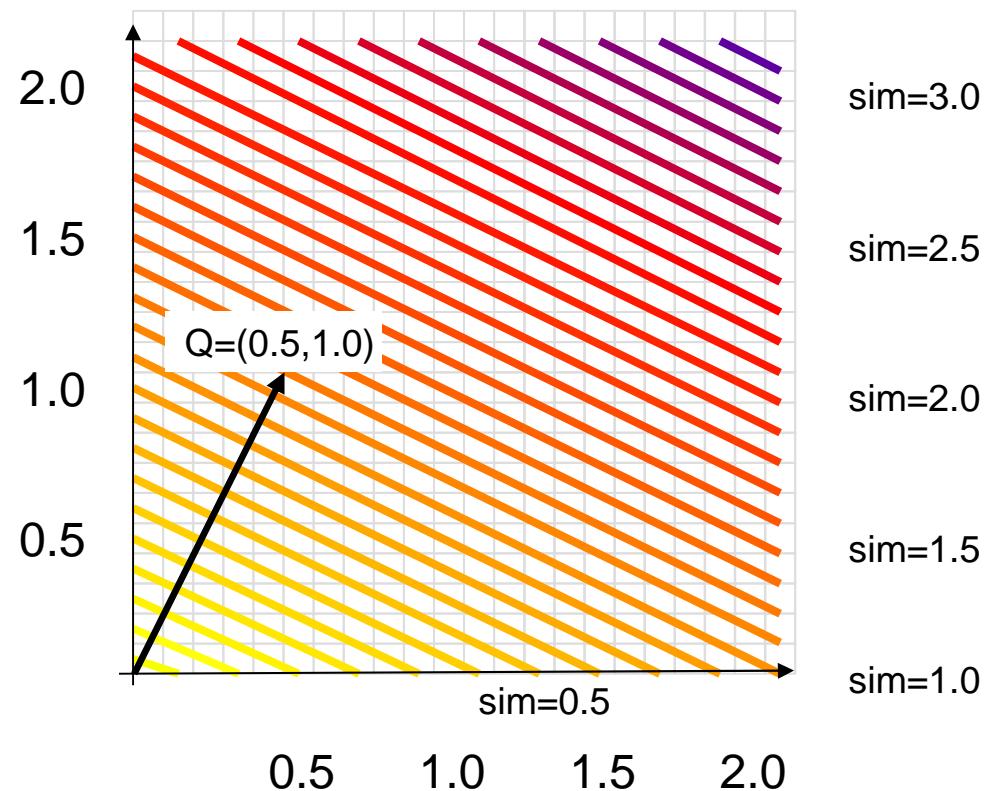
	d 1	d 5	sim (d1,d5)
Ätna	1	1	1 x 1
Vesuv	1	2	+ 1 x 2
Stromboli	1	3	+ 1 x 3
Feuer	1	4	+ 1 x 4
Wasser	1	5	+ 1 x 5
Lava	1	6	+ 1 x 6
			= 21



# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

## Skalarprodukt

### Iso-Konturen zum Referenzvektor $Q=(0.5,1.0)$

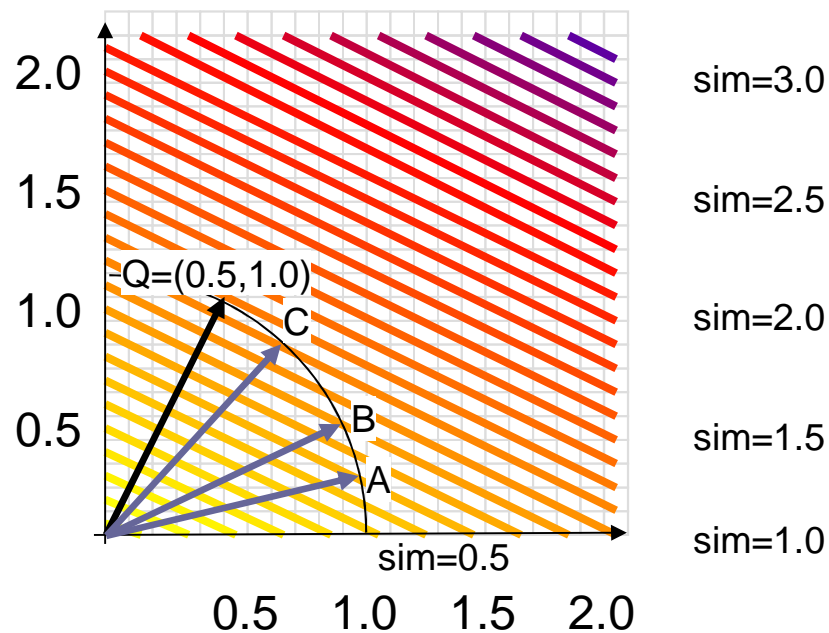


# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

## Skalarprodukt

### Parametermodifikation (1)

Veränderung des Winkels



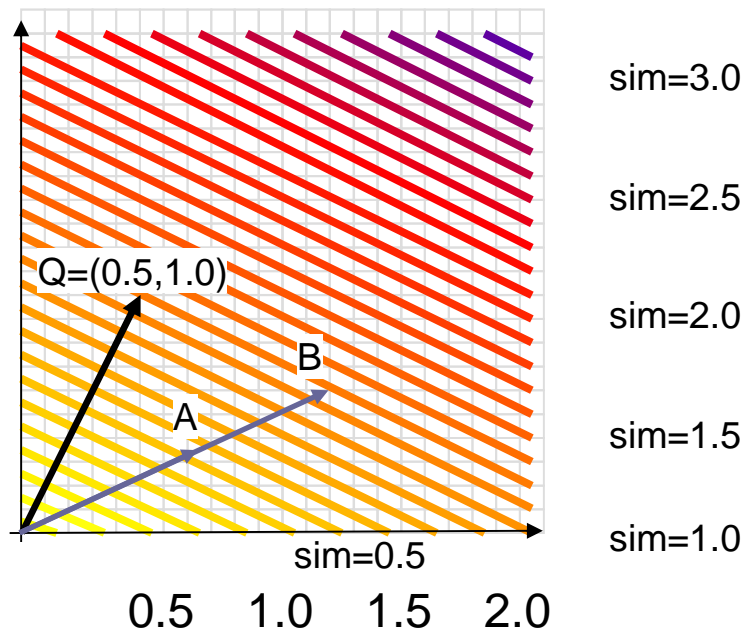
1. je kleiner der Winkel, desto größer die Ähnlichkeit – bei gleichlangen Vektoren
2. monoton – sofern Vektor gleichlang bleibt

# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

## Skalarprodukt

### Parametermodifikation (2)

Veränderung der Länge



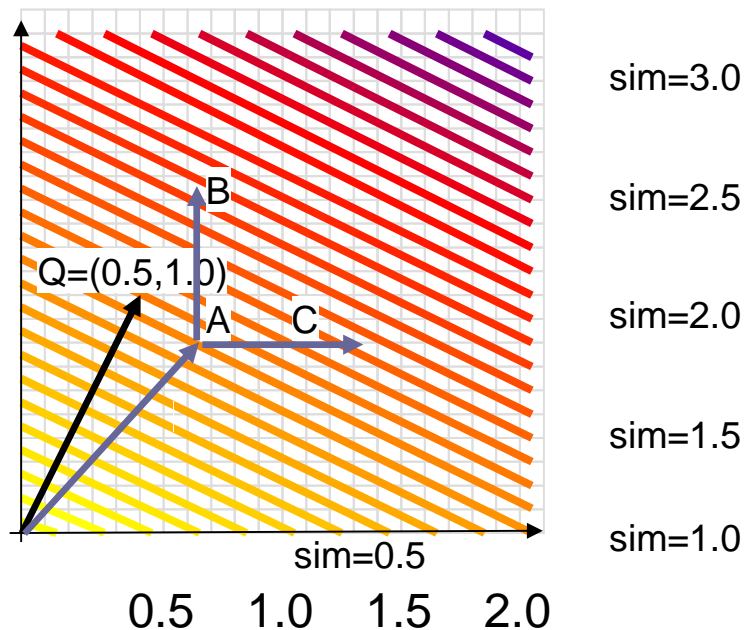
1. je länger der Vektor, desto größer die Ähnlichkeit
2. monoton

# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

## Skalarprodukt

### Parametermodifikation (3)

Addition von Komponenten



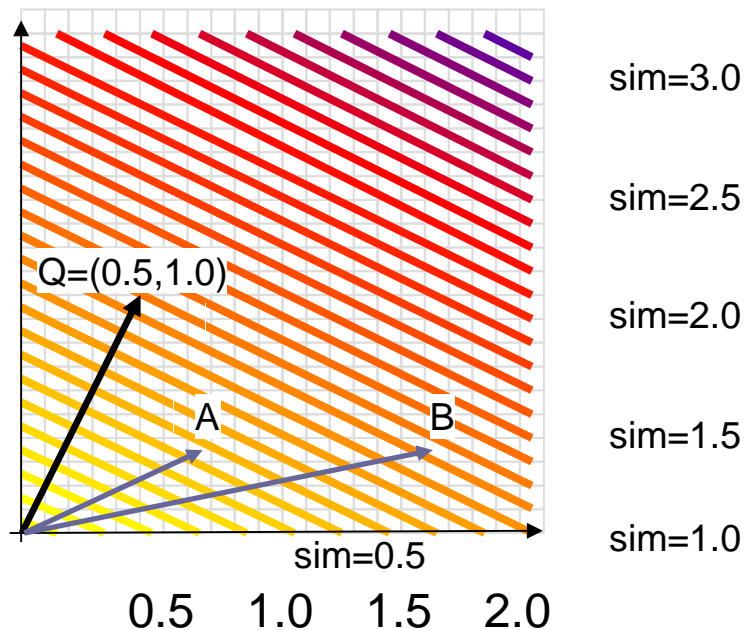
1. bei Addition jeder beliebigen Komponente bleibt Ähnlichkeit gleich oder steigt
2. monoton

# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

## Skalarprodukt

### Parametermodifikation (4)

Veränderung eines Einzelterms



1. unbegrenzter Einfluss einzelner Terme
2. monoton

# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

## Eigenschaften des Skalarprodukts

Eigenschaft	Verhalten	Bedeutung
Winkel	je kleiner der Winkel zwischen zwei Vektoren gleicher Euklidischer Länge, desto größer der Ähnlichkeitswert	Richtung (Thema) dominiert das Maß
Radius	längerer Vektor hat größeren oder gleichen Ähnlichkeitswert	„The more, the better“ kein Begriff einer adäquaten Tiefe
Komponenten	Verstärkung einzelner Komponenten: Ähnlichkeitswert größer, wenn Winkel zwischen den Vektoren kleiner wird, sonst gleich bleibend	
Einzelkomponenten	beliebig hoher Ähnlichkeitswert durch Veränderung eines einzelnen Wertes möglich	Einzelwerte können Ähnlichkeitswert dominieren; Objekte, die insgesamt unähnlich sind können einen sehr hohen Ähnlichkeitswert erhalten ▼
Wertebereich	bei nicht-negativen Werten $0 \leq sim \leq \infty$	es gibt kein Maximum, d.h. keinen Begriff einer idealen Bewertung ▼

# Eigenschaften des Skalarproduktes

## Beispiele

	d 1	d 2	d 3	d 4	d 5	d 6
Ätna	1	1	2	1	1	1
Vesuv	1	1	2	0	2	0
Stromboli	1	1	2	1	3	3
Feuer	1	1	2	0	4	0
Wasser	1	1	2	1	5	5
Lava	1	1	2	0	6	0

Einfache Übereinstimmung						
	d1	d2	d3	d4	d5	d6
d1	-	6.0	12.0	3.0	21.0	9.0
d2	6.0	-	12.0	3.0	21.0	9.0
d3	12.0	12.0	-	6.0	42.0	18.0
d4	3.0	3.0	6.0	-	9.0	9.0
d5	21.0	21.0	42.0	9.0	-	35.0
d6	9.0	9.0	18.0	9.0	35.0	-

- Erhöhung des Gewichtes eines beliebigen Terms hat proportionalen Effekt auf Ähnlichkeitswert des Dokuments<sup>1)</sup>
  - Beispiel:  $\text{sim}(d1, d4)$  vs.  $\text{sim}(d1, d6)$
- Beiträge verschiedener Terme sind voneinander unabhängig<sup>1)</sup>
  - hohe Werte für „Feuer“, „Wasser“, „Lava“ in d5 sorgen für hohe Ähnlichkeitswerte von d5
- absurde Ergebnisse bei Anwendung auf nicht-normalisierte Vektoren<sup>2)</sup>
  - Beispiel:  $\text{sim}(d1, d3) > \text{sim}(d1, d2)$ , obwohl d1 und d2 identisch sind

# Eigenschaften des Skalarproduktes

## Beispiele – Vorsicht!

	d 1	d 2	d 3	d 4	d 5	d 6
Ätna	1	1	2	1	1	1
Vesuv	1	1	2	0	2	0
Stromboli	1	1	2	1	3	3
Feuer	1	1	2	0	4	0
Wasser	1	1	2	1	5	5
Lava	1	1	2	0	6	0

Einfache Übereinstimmung						
	d1	d2	d3	d4	d5	d6
d1	-	6.0	12.0	3.0	21.0	9.0
d2	6.0	-	12.0	3.0	21.0	9.0
d3	12.0	12.0	-	6.0	42.0	18.0
d4	3.0	3.0	6.0	-	9.0	9.0
d5	21.0	21.0	42.0	9.0	-	35.0
d6	9.0	9.0	18.0	9.0	35.0	-

- Das Skalarprodukt wird im Information Retrieval zuweilen auch auf nicht-normalisierte Vektoren angewendet und „**einfache Methode**“ genannt
- Bei Anwendung auf nicht-normalisierte Vektoren ergeben sich aber absurde Ergebnisse:  
Beispiel:  $\text{sim}(d1, d3) > \text{sim}(d1, d2)$ , obwohl d1 und d2 identisch sind



# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

## Cosinus

$\text{sim}_{\cos}(\vec{X}, \vec{Y})$	Binäre Vektoren	Vektoren mit reellen Werten
Cosinus-Koeffizient <sub>allqVekt</sub>	$\frac{ X \cap Y }{\sqrt{ X  \times  Y }}$	$\frac{\sum_{k=1}^n x_k y_k}{\sqrt{\sum_{k=1}^n x_k^2} \cdot \sqrt{\sum_{k=1}^n y_k^2}}$

- Zähler: wie gut  $x_k$  und  $y_k$  korrelieren
- Nenner: Teilung durch (Euklidische) Länge der Vektoren

# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

## Eigenschaften des Cosinusmaßes

- Wertebereich von 1 bis -1
  - $\text{Cos}(0^\circ) = +1.0$  Vektoren zeigen in dieselbe Richtung
  - $\text{Cos}(90^\circ) = 0.0$  Vektoren orthogonal
  - $\text{Cos}(180^\circ) = -1.0$  Vektoren zeigen in entgegengesetzte Richtung
- Cosinus wirkt als normalisierender Korrelationskoeffizient
- Cosinus für normalisierte Vektoren entspricht Ähnlichkeit nach einfacher Methode (Skalarprodukt)

# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

## Eigenschaften des Cosinusmaßes

- Cosinusmaß ist identisch mit dem Skalarprodukt im Falle **normalisierter Vektoren**:

- normalisierter Vektor: Vektor mit Einheitslänge nach Euklidischer Norm

$$|\vec{x}| = \sqrt{\sum_{i=1}^n x_i^2} = 1$$

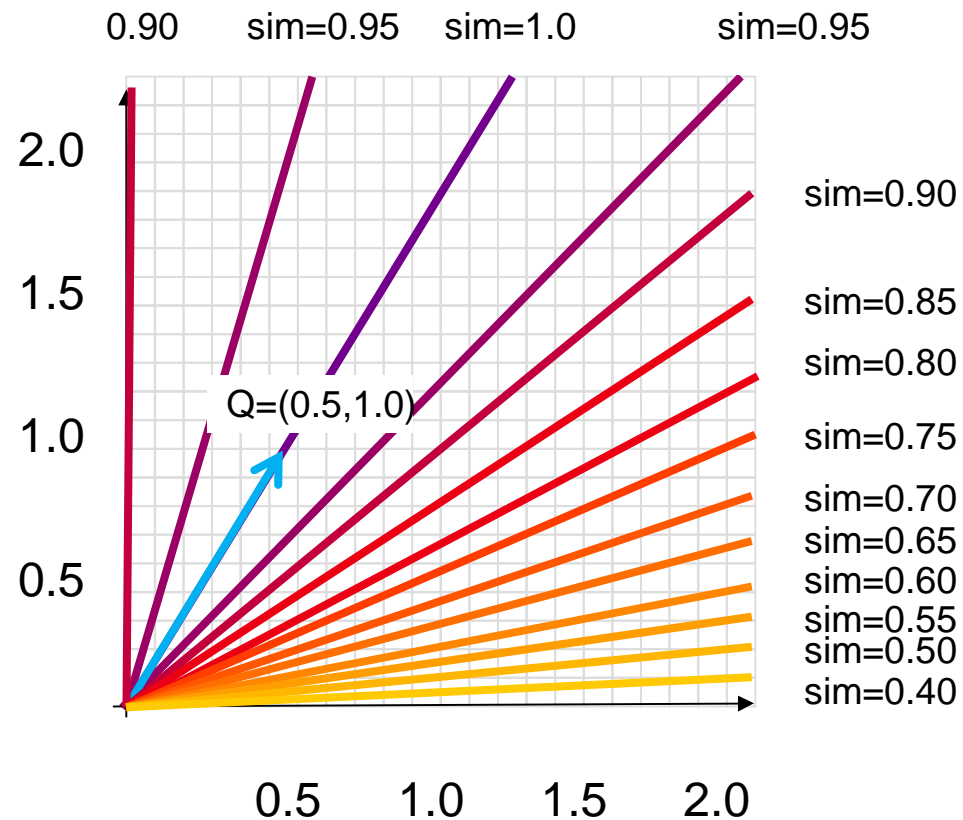
- für normalisierte Vektoren gilt:

$$\text{sim}(\vec{x}, \vec{y}) = \cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| \cdot |\vec{y}|} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \cdot \sqrt{\sum_{i=1}^n y_i^2}} = \vec{x} \cdot \vec{y}$$

# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

## Cosinus-Maß

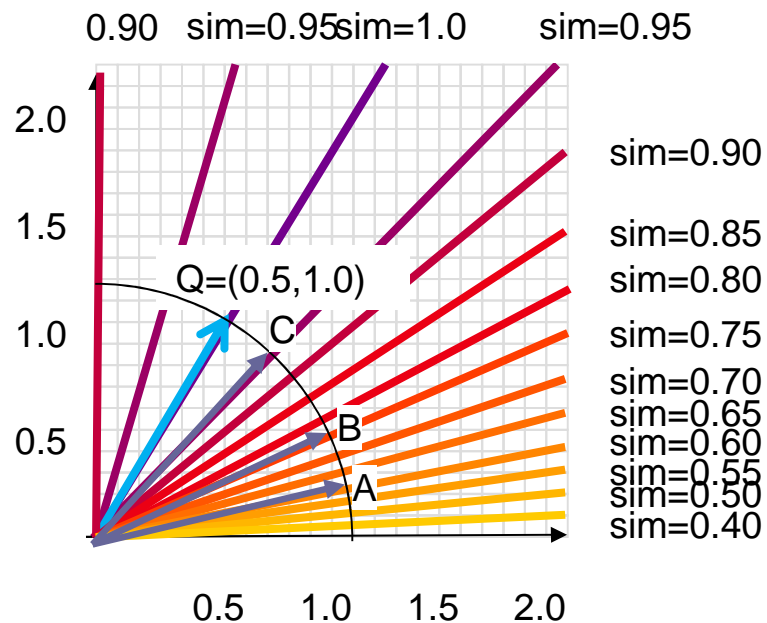
Iso-Konturen zum Referenzvektor  $Q=(0.5,1.0)$



# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

## Cosinusmaß

### Parametermodifikation (1)



### Veränderung des Winkels

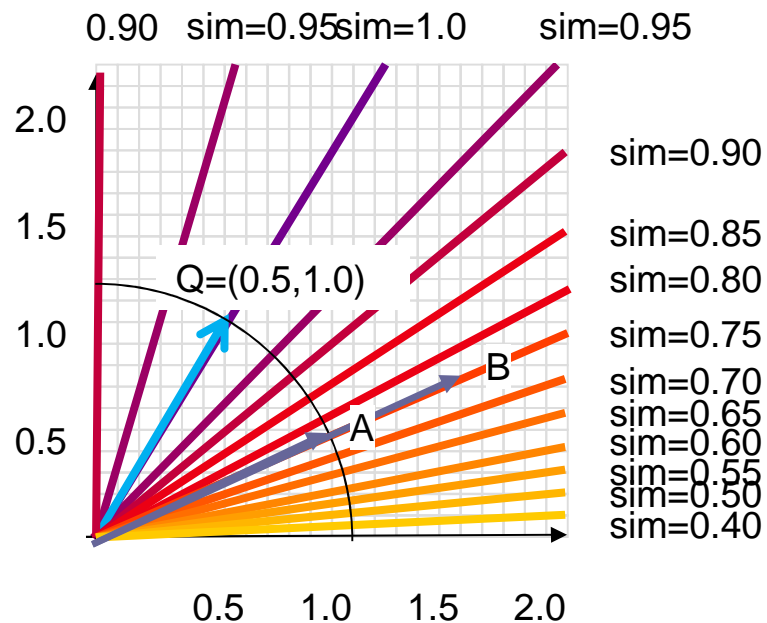
1. je kleiner der Winkel, desto größer die Ähnlichkeit – bei gleichlangen Vektoren
2. 1. gilt unabhängig von der Länge des Vektors
3. monoton – sofern Vektor gleichlang bleibt

# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

## Cosinusmaß

### Parametermodifikation (2)

Veränderung der Länge

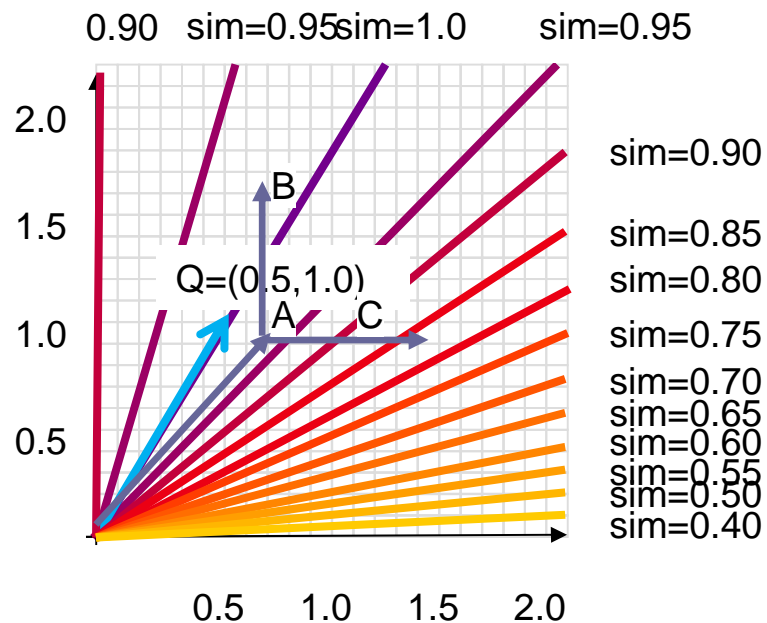


1. keine Veränderung des Maßes

# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

## Cosinusmaß

### Parametermodifikation (3)



### Addition von Komponenten

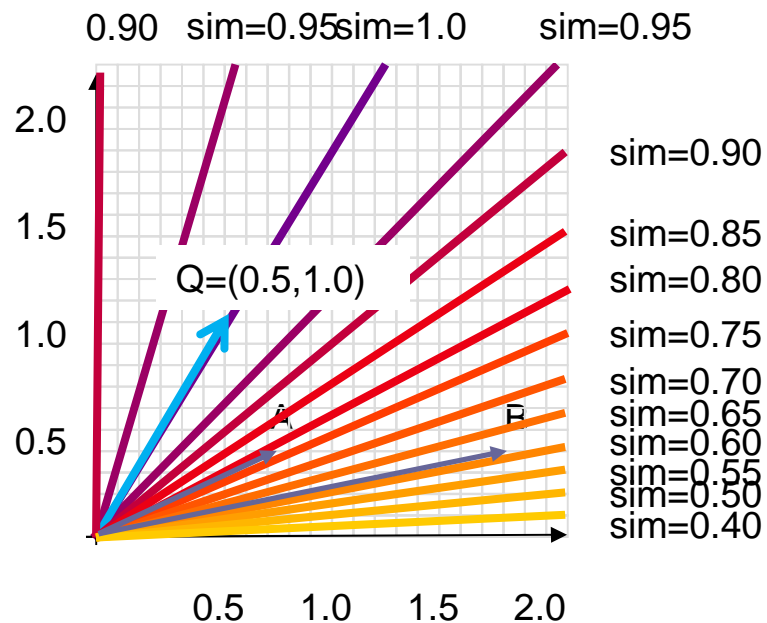
1. bei Addition jeder beliebigen Komponente ändert sich die Ähnlichkeit in Abhängigkeit von der Änderung des Winkels
2. monoton – abhängig von Veränderung des Winkels

# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

## Skalarprodukt

### Parametermodifikation (4)

Veränderung eines Einzelterms





1. abhängig von  
Veränderung des Winkels
2. monoton – abhängig von  
Veränderung des Winkels



# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

## Eigenschaften des Cosinusmaßes

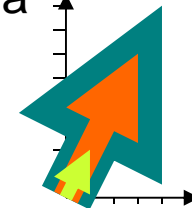
Eigenschaft	Verhalten	Bedeutung
Winkel	je kleiner der Winkel zwischen zwei Vektoren, desto größer der Ähnlichkeitswert	Richtung (Thema) dominiert das Cos-Maß
Radius	als Folge der Normalisierung keine Veränderung des Ähnlichkeitswertes bei Veränderung des Radius	Richtung (Thema) dominiert das Cos-Maß vollständig 
Komponenten	Verstärkung einzelner Komponenten: Ähnlichkeitswert - wird größer, wenn dadurch der Winkel zwischen den Vektoren verkleinert wird - wird kleiner, wenn dadurch der Winkel zwischen den Vektoren vergrößert wird	abhängig von Veränderung des Winkels
Einzelkomponenten	Ähnlichkeitsmaß bestimmt durch Ähnlichkeit der Termgewichtsproportionen (Verhältnis der Werte in den einzelnen Vektoren)	abhängig von Veränderung des Winkels
Wertebereich	bei nicht-negativen Werten $0 \leq sim \leq 1$	es gibt ein Maximum  d.h. einen Idealwert

# Eigenschaften des Cosinus-Maßes

## Beispiele

	d 1	d 2	d 3	d 4	d 5	d 6	Cosinus						
Ätna	1	1	2	1	1	1		d1	d2	d3	d4	d5	d6
Vesuv	1	1	2	0	2	0	d1	-	1.000	1.000	0.707	0.898	0.621
Stromboli	1	1	2	1	3	3	d2	1.000	-	1.000	0.707	0.890	0.621
Feuer	1	1	2	0	4	0	d3	1.000	1.000	-	0.707	0.898	0.621
Wasser	1	1	2	1	5	5	d4	0.707	0.707	0.707	-	0.544	0.878
Lava	1	1	2	0	6	0	d5	0.898	0.898	0.898	0.544	-	0.620
							d6	0.621	0.621	0.621	0.878	0.620	-

- **Ähnlichkeitswert eines Dokuments** wird allein durch sein „Thema“ (Relation der Terme innerhalb des Dokuments) bestimmt<sup>1)</sup>
  - Beispiel:  $\text{sim}(d1, d2) = \text{sim}(d1, d3)$
- **Termgewichtsbeziehungen zwischen Dokumenten** werden möglicherweise ignoriert<sup>1)</sup>
  - Beispiel:  $\text{sim}(d5, d1) > \text{sim}(d5, d6)$
- **Nullwerte** haben große Auswirkung auf das Ergebnis<sup>1)</sup>
  - Beispiel:  $\text{sim}(d1, d5) > \text{sim}(d1, d4)$



# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

## Dice-Koeffizient

	Binäre Vektoren	Vektoren mit reellen Werten
Dice-Koeffizient	$\frac{2   X \cap Y  }{  X   +   Y  }$	$\frac{2 \sum_{k=1}^n (weight_{xk} \cdot weight_{yk})}{\sum_{k=1}^n weight_{xk} + \sum_{k=1}^n weight_{yk}}$

- Einbeziehen des Anteils von gemeinsamen Einträgen:  
Summe aus gemeinsamen Einträgen, die  $\neq 0$  sind  
relativ zu Summe aus allen einträgen, die  $\neq 0$  sind
- Multiplikation mit 2, um bei binären Vektoren einen Wertebereich zwischen 0 und 1 zu erhalten
- Eigenschaften (Reaktion auf Länge, Reaktion auf Winkel)  
variieren bei reellen Werten in Abhängigkeit von der Relation der Länge der beiden Vektoren

# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

## Dice-Koeffizient

	Binäre Vektoren	Vektoren mit reellen Werten
Dice-Koeffizient	$\frac{2   X \cap Y  }{  X   +   Y  }$	$\frac{2 \sum_{k=1}^n (weight_{xk} \cdot weight_{yk})}{\sum_{k=1}^n weight_{xk} + \sum_{k=1}^n weight_{yk}}$

$|X|$  Anzahl der Nicht-Null-Werte in den binären Vektoren

Beispiel:  $|X| = 1$  Vektor mit 1 Nicht-Null-Wert  
 $|Y| = 1000$  Vektor mit 1000 Nicht-Null-Werten  
 $|X| \cap |Y| = 1$  1 gemeinsamer Eintrag

Berechnung nach der Formel  
für binäre Werte

$$\frac{2 \cdot 1}{1 + 1000} \approx 0.002$$

Berechnung nach der Formel  
für reelle Werte

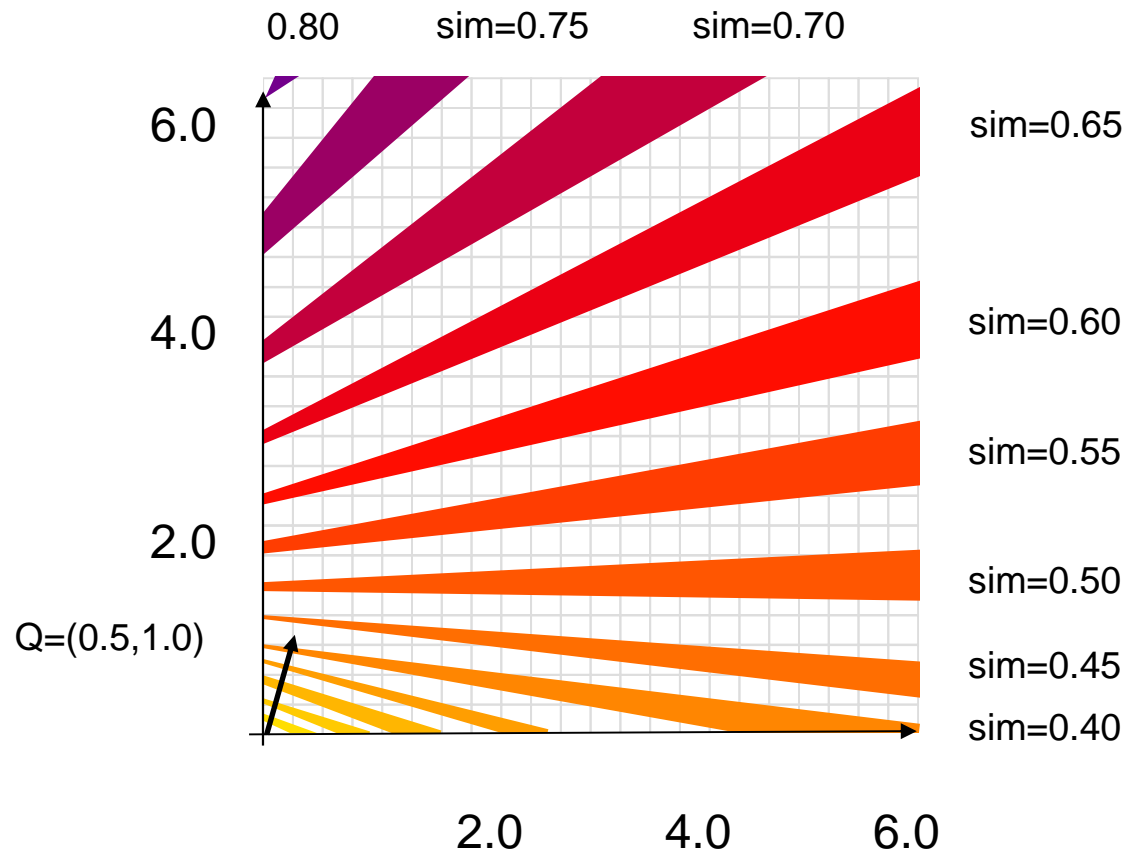
(seien 0 und 1 reelle Gewichte)

$$\frac{2 \cdot (1 \cdot 1 + 999(1 \cdot 0))}{1 + 1000} = \frac{2 \cdot 1}{1 + 1000} \approx 0.002$$

# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

## Dice-Koeffizient

Iso-Konturen zum Referenzvektor  $Q=(0.5,1.0)$



# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

## Dice-Koeffizient

	Binäre Vektoren	Vektoren mit reellen Werten
Dice-Koeffizient	$\frac{2   X \cap Y  }{  X   +   Y  }$	$\frac{2 \sum_{k=1}^n (weight_{xk} \cdot weight_{yk})}{\sum_{k=1}^n weight_{xk} + \sum_{k=1}^n weight_{yk}}$

- Nenner der Formel ist additiv (Addition der City-Block-Länge der beiden Vektoren – L1)
- **Isokonturlinien verändern sich** in Abhängigkeit der relativen L1-Länge von Anfrage- und Dokument-Vektor
  - beliebig langer Anfragevektor im Verhältnis zum Dokumentvektor: Einfluss der Länge des Dokumentvektors wird beliebig klein → Dice-Maß wird dem **Skalarprodukt ähnlich**
  - beliebig langer Dokumentvektor im Verhältnis zum Anfragevektor: Einfluss der Länge des Anfragevektors wird beliebig klein → Dice-Maß wird dem **Pseudo-Cosinus-Maß ähnlich**

# Ähnlichkeitsmaße im Information Retrieval

## Cosinus- vs. Dice-Koeffizient

- Cosinus wie Dice-Koeffizient für Vektoren mit derselben Anzahl von Nicht-Null-Werten
- Cosinus höhere Werte als Dice-Koeffizient, wenn Anzahl der Nicht-Null-Werte in den betrachteten Vektoren sehr verschieden ist
- Beispiel:
  - Vektor 1: 1 Nicht-Null-Eintrag
  - Vektor 2: 1000 Nicht-Null-Einträge
  - 1 gemeinsamer Eintrag

$$Dice \frac{2 \times 1}{1 + 1000} \approx 0.002$$

$$Cosinus \frac{1}{\sqrt{1000 \times 1}} \approx 0.03$$

---

(Manning/Schütze, 2000, 300/301)

# Ähnlichkeitsmaße im Information Retrieval

## Jaccard-Koeffizient

	Binäre Vektoren	Vektoren mit reellen Werten
Jaccard (oder Tanimoto)-Koeffizient	$\frac{ X \cap Y }{ X \cup Y }$	$\frac{\sum_{k=1}^n (weight_{xk} \cdot weight_{yk})}{\sum_{k=1}^n weight_{xk} + \sum_{k=1}^n weight_{yk} - \sum_{k=1}^n (weight_{xk} \cdot weight_{yk})}$ <p>(Ferber, 2003)</p>

- bestraft Vorhandensein einer kleinen Anzahl gemeinsamer Einträge stärker als Dice-Koeffizient  
(je weniger gemeinsame Einträge, desto größer der Nenner, desto kleiner der Wert des Bruches)
- Beispiel: 2 Vektoren, 10 Nicht-Null-Einträge, 1 gemeinsamer Eintrag

$$Dice \frac{2 \times 1}{10 + 10} = 0.1$$

$$Jaccard \frac{1}{10 + 10 - 1} \approx 0.05$$

(Manning/Schütze, 2000)



# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

## Overlap-Koeffizient

	Binäre Vektoren	Vektoren mit reellen Werten
Overlap-Koeffizient	$\frac{ X \cap Y }{\min( X ,  Y )}$	$\frac{\sum_{k=1}^n \min(\text{weight}_{xk}, \text{weight}_{yk})}{\min(\sum_{k=1}^n \text{weight}_{xk}, \sum_{k=1}^n \text{weight}_{yk})}$

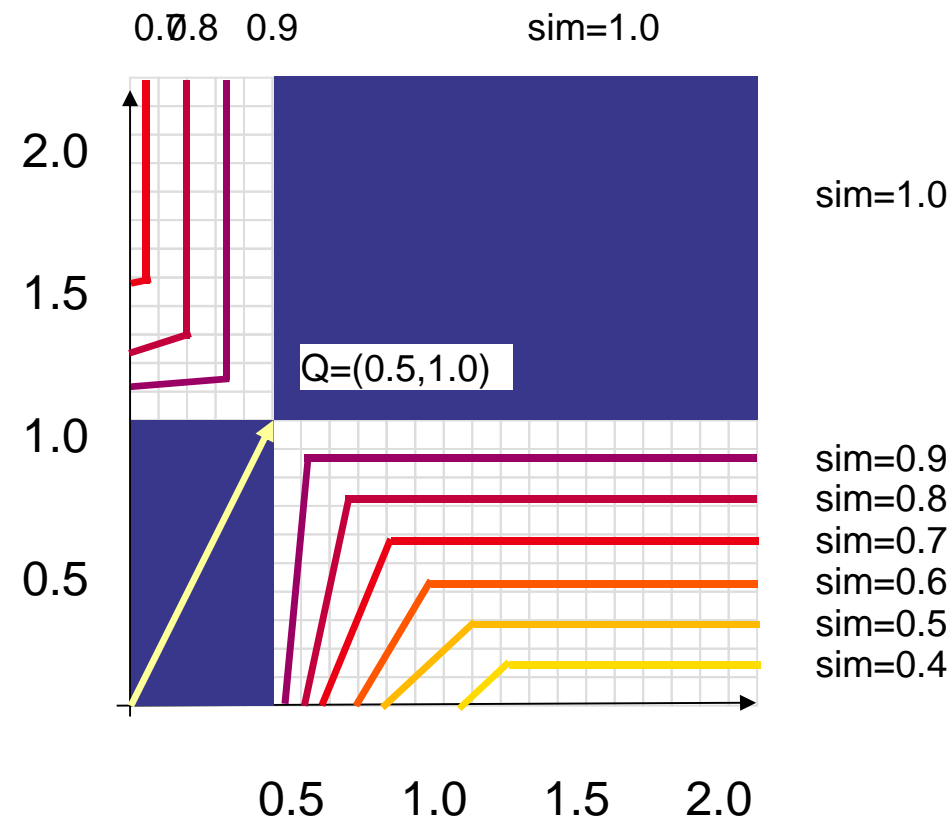
(Ferber, 2003)

- Maß für Inklusion
- erreicht Wert von 1.0 (binäre Vektoren), wenn jede Dimension mit Nicht-Null-Wert in Vektor X auch in Vektor Y Nicht-Null-Wert hat, und umgekehrt

# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

## Overlap-Maß

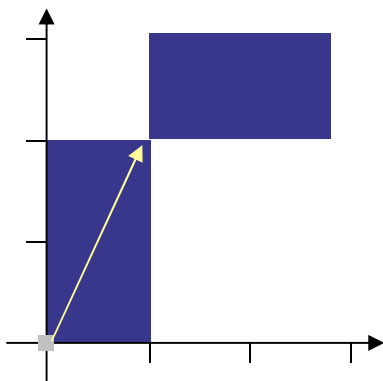
Iso-Konturen zum Referenzvektor  $Q=(0.5,1.0)$



# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

## Overlap-Maß

- durch die Minima, die in der Formel auftreten, treten komplexere Gebilde gleicher Ähnlichkeit auf, die durch die Fallunterscheidung bei der Minimumbildung verursacht werden<sup>1)</sup>
- zwei Regionen maximaler Ähnlichkeit:
  - Vektor<sub>1</sub> in allen Dimensionen < Vektor<sub>2</sub>
  - Vektor<sub>1</sub> in allen Dimensionen > Vektor<sub>2</sub><sup>2)</sup>
- Veränderungen sind nicht monoton



<sup>1)</sup>Ferber, 2003

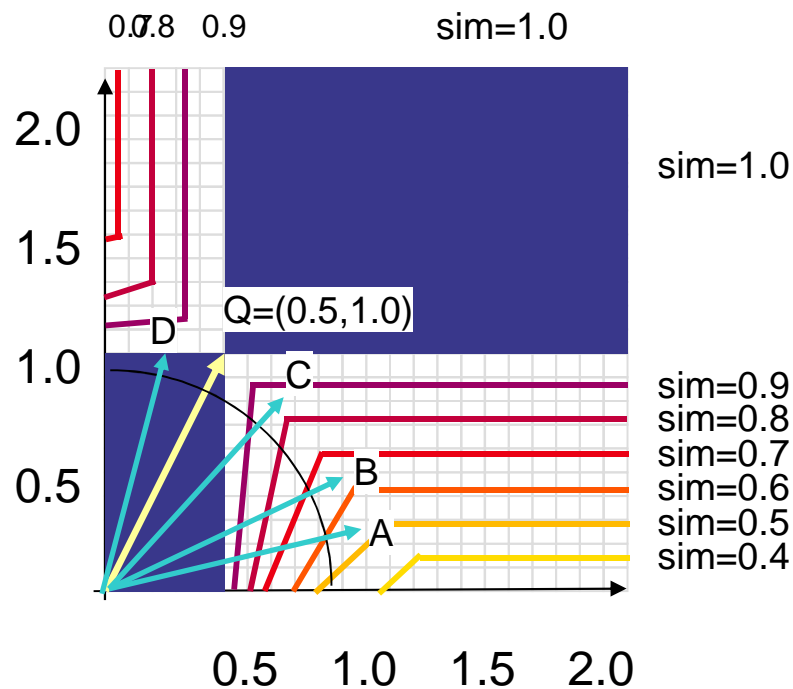
<sup>2)</sup>Jones/Furnas, 1987

# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

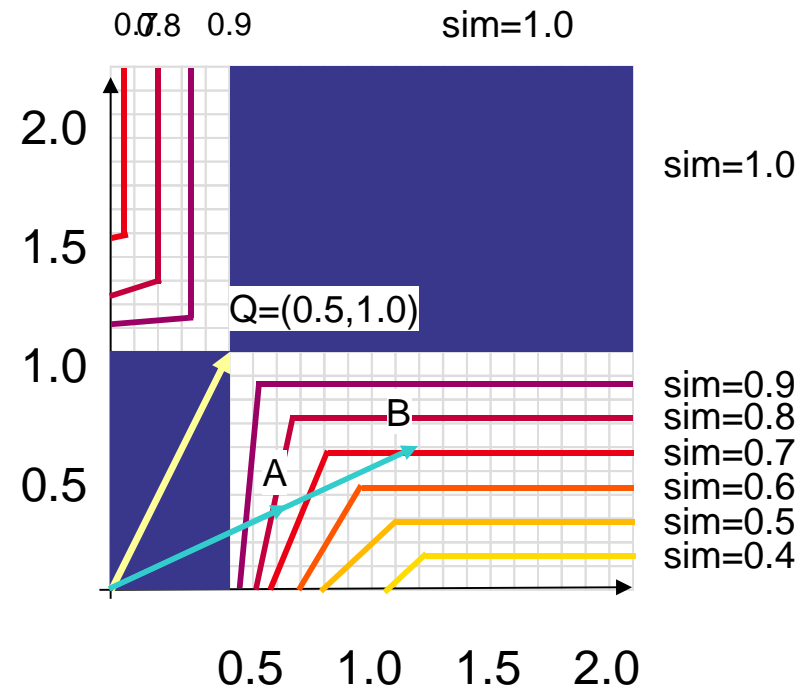
## Overlapmaß

### Parametermodifikation (1)

Veränderung des Winkels



Veränderung der Länge

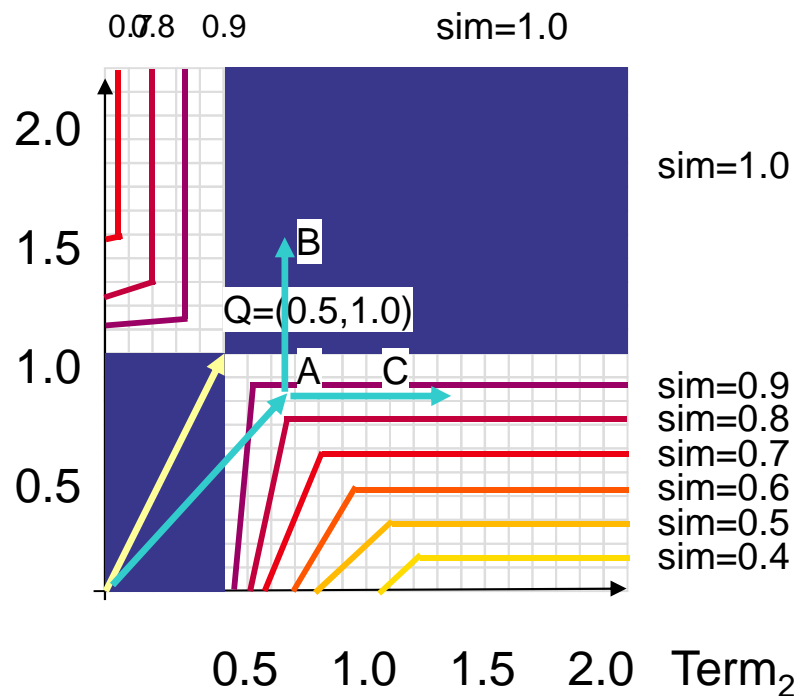


# Eine Analyse der Ähnlichkeitsmaße (Jones/Furnas)

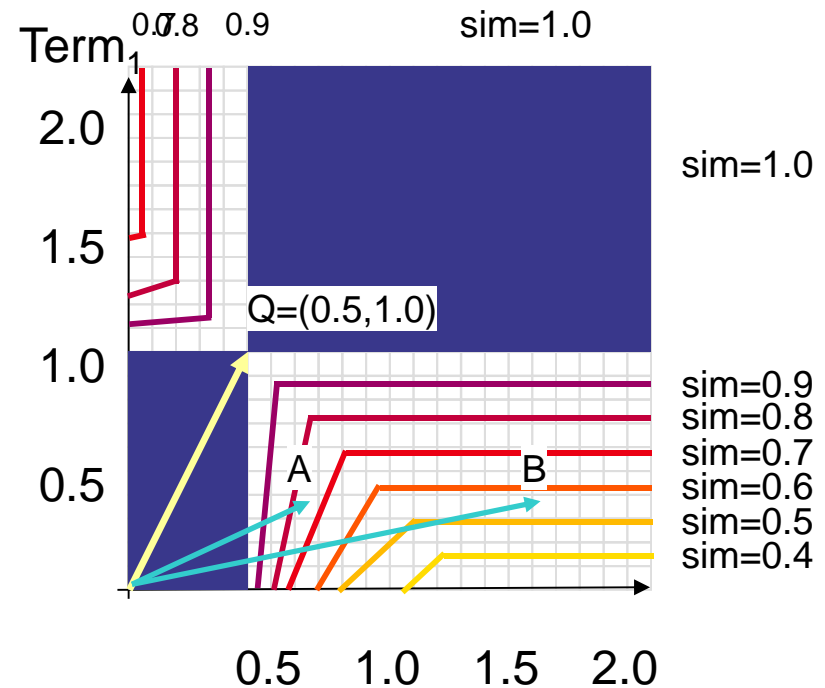
## Overlapmaß

### Parametermodifikation (2)

Addition von Komponenten



Veränderung eines Einzelterms



# Eigenschaften des Overlap-Koeffizienten

## Beispiele

	d 1	d 2	d 3	d 4	d 5	d 6	Overlap						
Ätna	1	1	2	1	1	1		d1	d2	d3	d4	d5	d6
Vesuv	1	1	2	0	2	0	d1	-	1.000	1.000	1.000	1.000	0.500
Stromboli	1	1	2	1	3	3	d2	1.000	-	1.000	1.000	1.000	0.500
Feuer	1	1	2	0	4	0	d3	1.000	1.000	-	1.000	0.916	0.555
Wasser	1	1	2	1	5	5	d4	1.000	1.000	1.000	-	1.000	1.000
Lava	1	1	2	0	6	0	d5	1.000	1.000	0.916	1.000	-	1.000
							d6	0.500	0.500	0.555	1.000	1.000	-

- Maxima (1.0) bei allen Fällen
  - $V_1 < V_2$  in allen Dimensionen und
  - $V_1 > V_2$  in allen Dimensionen
- favorisiert Vektoren, die entweder sehr lang oder sehr kurz sind
- allgemeine Unempfindlichkeit gegenüber objekt-internen und objekt-übergreifenden Termgewichtsbeziehungen

Jones/Furnas, 1987

# Distanzmaße

## Euklidische Distanz

$$|\vec{x} - \vec{y}| = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

liefert für normalisierte Vektoren dasselbe Ranking wie Cosinus-Maß, wie die folgende Ableitung zeigt

$$\begin{aligned} \left(|\vec{x} - \vec{y}|\right)^2 &= \sum_{i=1}^n (x_i - y_i)^2 \\ &= \sum_{i=1}^n x_i^2 - 2 \sum_{i=1}^n x_i y_i + \sum_{i=1}^n y_i^2 \\ &= 1 - 2 \sum_{i=1}^n x_i y_i + 1 \\ &= 2(1 - \vec{x} \cdot \vec{y}) \end{aligned} \quad (\text{Manning/Schütze, 2000})$$

# Inhalt

- Einführung
  - Ähnlichkeitsmaß
  - Ähnlichkeitsbetrachtungen
- Gebräuchliche Ähnlichkeitsmaße im Information Retrieval
  - Korrelationsmaße: Einfache Methode, Cosinus, Dice-Koeffizient, Jaccard-Koeffizient, Overlap-Koeffizient
  - Distanzmaße: Euklidische Distanz
- Eine Analyse der Ähnlichkeitsmaße von Jones/Furnas (1987)
- Beispiel 1: Berechnung der Ähnlichkeitsmaße für sechs Dokumentvektoren
- Beispiel 2: Bestimmung der Ähnlichkeit von Nomina auf der Basis von Prädikat-Objekt-Kookkurrenz-Paaren



	d 1	d 2	d 3	d 4	d 5	d 6
Ätna	1	1	2	1	1	1
Vesuv	1	1	2	0	2	0
Stromboli	1	1	2	1	3	3
Feuer	1	1	2	0	4	0
Wasser	1	1	2	1	5	5
Lava	1	1	2	0	6	0

Einfache Übereinstimmung						
	d1	d2	d3	d4	d5	d6
d1	-	6.0	12.0	3.0	21.0	9.0
d2	6.0	-	12.0	3.0	21.0	9.0
d3	12.0	12.0	-	6.0	42.0	18.0
d4	3.0	3.0	6.0	-	9.0	9.0
d5	21.0	21.0	42.0	9.0	-	35.0
d6	9.0	9.0	18.0	9.0	35.0	-

Dice						
	d1	d2	d3	d4	d5	d6
d1	-	1.000	1.333	0.666	1.555	1.200
d2	1.000	-	1.333	0.666	1.555	1.200
d3	1.333	1.333	-	0.800	2.545	1.714
d4	0.666	0.666	0.800	-	0.750	1.500
d5	1.555	1.555	2.545	0.750	-	2.333
d6	1.200	1.200	1.714	1.500	2.333	-

Cosinus						
	d1	d2	d3	d4	d5	d6
d1	-	1.000	1.000	0.707	0.898	0.621
d2	1.000	-	1.000	0.707	0.890	0.621
d3	1.000	1.000	-	0.707	0.898	0.621
d4	0.707	0.707	0.707	-	0.544	0.878
d5	0.898	0.898	0.898	0.544	-	0.620
d6	0.621	0.621	0.621	0.878	0.620	-

Jaccard						
	d1	d2	d3	d4	d5	d6
d1	-	1.000	2.000	0.500	3.500	1.500
d2	1.000	-	2.000	0.500	3.500	1.500
d3	2.000	2.000	-	0.666	-4.66	6.000
d4	0.500	0.500	0.666	-	0.600	3.000
d5	3.500	3.500	-4.66	0.600	-	-7.00
d6	1.500	1.500	6.000	3.000	-7.00	-

Overlap						
	d1	d2	d3	d4	d5	d6
d1	-	1.000	1.000	1.000	1.000	0.500
d2	1.000	-	1.000	1.000	1.000	0.500
d3	1.000	1.000	-	1.000	0.916	0.555
d4	1.000	1.000	1.000	-	1.000	1.000
d5	1.000	1.000	0.916	1.000	-	1.000
d6	0.500	0.500	0.555	1.000	1.000	-

# Auswahlkriterien von Ähnlichkeitsmaßen für Vektoren

- Mathematische Eigenschaften der Ähnlichkeitsmaße
- Empirische Evaluierung
  - Art der Datenbasis
  - Benutzungssituationen
  - Informationsbedarf
  - Kriterien der Erstellung der Dokumentrepräsentationen
- unklar, ob und wie mathematische Eigenschaften mit pragmatischen Faktoren zusammenpassen

---

Jones/Furnas, 1987: 420

# Inhalt

- Einführung
  - Ähnlichkeitsmaß
  - Ähnlichkeitsbetrachtungen
- Gebräuchliche Ähnlichkeitsmaße im Information Retrieval
  - Korrelationsmaße: Einfache Methode, Cosinus, Dice-Koeffizient, Jaccard-Koeffizient, Overlap-Koeffizient
  - Distanzmaße: Euklidische Distanz
- Eine Analyse der Ähnlichkeitsmaße von Jones/Furnas (1987)
- Beispiel 1: Berechnung der Ähnlichkeitsmaße für sechs Dokumentvektoren
- Beispiel 2: Bestimmung der Ähnlichkeit von Nomina auf der Basis von Prädikat-Objekt-Kookkurrenz-Paaren

## Ein Beispiel

**Distributional similarity measures:** The program illustrates 11 different distributional similarity measures:

Cosine

Jaccard coefficient

Dice coefficient

Overlap coefficient

L1 distance (City block distance)

Euclidean distance (L2 distance)\* (\* applied to non-normalized vectors)

Hellinger distance

Information Radius (Jensen-Shannon divergence)

Skew divergence\*\* (\*\*  $\alpha = 0.001$ )

Confusion Probability

Lin's Similarity Measure

Based on distributional data from BNC (**predicate-object co-occurrence pairs**), for an input noun, the program retrieves 30 most similar ones. It also describes the number of non-zero features for the target noun as well as its frequency rank in the dataset.

## Beispielsergebnisse für „water“

Cosinus	Dice	Jaccard	Overlap
whisky	air	air	part
milk	part	part	line
wine	line	line	body
brandy	place	place	room
tea	house	house	air
sherry	room	room	car
coffee	thing	thing	house
champagne	hand	hand	hand
liquid	area	area	face
gin	car	car	area
cup	body	body	arm
juice	arm	arm	head
encourager	paper	paper	world
le	food	food	case
tonic	box	box	back
refill	number	number	home

## Beispielsergebnisse für „water“

Euklid.Dist.	L1-Distance (city block)	Information Radius	Lin Similarity
man thing body number kind way woman side line work place area person people sort part	cup glass bottle line milk part air wine place thing house body coffee room area tea	milk wine cup coffee bottle tea air glass line juice place part house whiskey oil river	air river blood pool bath sea wine food room place arm pocket bottle tea mouth amount

# Vielen Dank

Für das Aufspüren von Fehlern in früheren Versionen und für Verbesserungsvorschläge danke ich

Nicola Kaiser, Sebastian Kreß, Philipp Scheffzek, Wolodja Wentland

Für Hinweise zum Overlap-Koeffizienten danke ich

Reginald Ferber

# Literatur

- Ferber, Reginald (2003). *Information Retrieval. Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web*. Heidelberg: dpunkt-Verlag. <http://information-retrieval.de/irb/ir.html>
- Jones, William P. und George W. Furnas (1987). Pictures of Relevance: A Geometric Analysis of Similarity Measures. In: *Journal of the American Society for Information Science*. 38 (6), S. 420-442.
- Manning, Christopher; Schütze, Hinrich (2000): *Foundations of Statistical Natural Language Processing*. Cambridge, Mass.: MIT Press.
- Pekar, Viktor. *Distributional Similarity Measures online demo*. <http://clg.wlv.ac.uk/demos/similarity/index.html>. (Calculates distributionally similar words according distributional similarity measures for nouns in the BNC.)
- Rijsbergen, C. J. (1979). *Information Retrieval*. Sec. Ed. London: Butterworths.



# Copyright

- © Karin Haenelt, 2000,2006,2007, 2012  
All rights reserved. The German [Urheberrecht](#) (esp. § 2, § 13, § 63 , etc.). shall be applied to these slides. In accordance with these laws these slides are a publication which may be quoted and used for non-commercial purposes, if the bibliographic data is included as described below.
  - Please quote correctly.
    - If you use the presentation or parts of it for educational and scientific purposes, please include the bibliographic data (author, title, date, page, URL) in your publication (book, paper, course slides, etc.).
    - please add a bibliographic reference to copies and quotations
  - *Deletion or omission of the footer (with name, data and copyright sign) is not permitted if slides are copied*
  - *Bibliographic data. Karin Haenelt, Ähnlichkeitsmaße für Vektoren. Kursfolien 25.10.2012 (1. Fassung 15.11.2000) + URL*
- *For commercial use:* In case you are interested in commercial use please contact the author.
- Court of Jurisdiction is Darmstadt, Germany

---

versions: Ähnlichkeitsmaße 25.10.2012, 28.10.2007, 21.10.2007, 09.02.2007, 26.11.2006, 22.11.2004;in: Clustering 15.11.2000