

Multimedia Retrieval im WS 2011/2012

5. Distanzfunktionen

Prof. Dr.-Ing. Marcin Grzegorzek
Juniorprofessur für Mustererkennung
Institut für Bildinformatik im Department ETI
Fakultät IV der Universität Siegen

5. und 6. Dezember 2011



5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

1. Einführung

1.1 Grundlegende Begriffe

1.2 Suche in einem MMDBS

1.3 MMDBMS-Anwendungen

11.10.2011

2. Prinzipien des Information Retrieval

2.1 Einführung

2.2 Information-Retrieval-Modelle

2.3 Relevance Feedback

2.4 Bewertung von Retrieval-Systemen

17.10.2011

2.5 Nutzerprofile

3. Prinzipien des Multimedia Retrieval

3.1 Besonderheiten der Verwaltung und des Retrievals

3.2 Ablauf des Multimedia-Information-Retrievals

3.3 Daten eines Multimedia-Retrieval-Systems 24.10.2011

3.4 Feature

3.5 Eignung verschiedener Retrieval-Modelle

3.6 Multimedia-Ähnlichkeitsmodell 25.10.2011

4. Feature-Transformationsverfahren

4.1 Diskrete Fourier-Transformation 08.11.2011

4.2 Diskrete Wavelet-Transformation 14.11.2011

4.3 Karhunen-Loeve-Transformation 21.11.2011

4.4 Latent Semantic Indexing und Singulärwertzerlegung 22.11.2011

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Inhalte und Termine

5. Distanzfunktionen

5.1 Eigenschaften und Klassifikation

5.2 Distanzfunktionen auf Punkten

5.3 Distanzfunktionen auf Binärdaten

05.12.2011

5.4 Distanzfunktionen auf Sequenzen

5.5 Distanzfunktionen auf allgemeinen Mengen

06.12.2011

6. Ähnlichkeitsmaße

6.1 Einführung

6.2 Distanz versus Ähnlichkeit

6.3 Grenzen von Ähnlichkeitsmaßen

6.4 Konkrete Ähnlichkeitsmaße

6.5 Aggregation von Ähnlichkeitswerten

6.6 Umwandlung von Distanzen in Ähnlichkeitswerte und Normierung

6.7 Partielle Ähnlichkeit

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

7. Effiziente Algorithmen und Datenstrukturen

7.1 Hochdimensionale Indexstrukturen

7.2 Algorithmen zur Aggregation von Ähnlichkeitswerten

8. Anfragebehandlung

8.1 Einführung

8.2 Konzepte der Anfragebehandlung

8.3 Datenbankmodell

8.4 Sprachen

9. Zusammenfassung

Überblick

5.1
Eigenschaften

5.1 Eigenschaften und Klassifikation

5.2 DF auf
Punkten

5.2 Distanzfunktionen auf Punkten

5.3 DF auf
Binärdaten

5.3 Distanzfunktionen auf Binärdaten

5.4 DF auf
Sequenzen

5.4 Distanzfunktionen auf Sequenzen

5.5 DF auf
allgemeinen
Mengen

5.5 Distanzfunktionen auf allgemeinen Mengen

Überblick

5.1
Eigenschaften

5.1 Eigenschaften und Klassifikation

5.2 DF auf
Punkten

5.2 Distanzfunktionen auf Punkten

5.3 DF auf
Binärdaten

5.3 Distanzfunktionen auf Binärdaten

5.4 DF auf
Sequenzen

5.4 Distanzfunktionen auf Sequenzen

5.5 DF auf
allgemeinen
Mengen

5.5 Distanzfunktionen auf allgemeinen Mengen

Eigenschaften und Klassifikation

5.1 Eigenschaften

5.2 DF auf Punkten

5.3 DF auf Binärdaten

5.4 DF auf Sequenzen

5.5 DF auf allgemeinen Mengen

- Abbildung Feature-Werte zweier Medienobjekte auf **nichtnegative, reelle Zahl**
- Distanzwert 0 bedeutet maximale Ähnlichkeit
- **Invarianz** einer Distanzfunktion
→ Unabhängigkeit bzgl. Operation g :
$$d(g(o_1), g(o_2)) = d(o_1, o_2)$$

Operation g ist etwa

- Translation
- Skalierung
- Rotation

Formale Eigenschaften einer Distanzfunktion

5.1 Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

binäre Funktion $d(o_1, o_2)$ mit $d : O \times O \longrightarrow \mathbb{R}_0^+$ und

- **Selbstidentität** (Si): $\forall o \in O : d(o, o) = 0$
- **Positivität** (Pos): $\forall o_1 \neq o_2 \in O : d(o_1, o_2) > 0$
- **Symmetrie** (Sym): $\forall o_1, o_2 \in O : d(o_1, o_2) = d(o_2, o_1)$
- **Dreiecksungleichung** (Dreieck):

$$\forall o_1, o_2, o_3 \in O : d(o_1, o_3) \leq d(o_1, o_2) + d(o_2, o_3)$$

Klassifikation anhand Erfüllung der Eigenschaften

5.1 Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Klasse	Si	Pos	Sym	Dreieck
Distanzfunktion	✓	✓	✓	✓
Pseudo-Distanzfunktion	✓	–	✓	✓
Semi-Distanzfunktion	✓	✓	✓	–
Semi-Pseudo-Distanzfunktion	✓	–	✓	–

Beispiele von Distanzfunktionen

5.1 Eigenschaften

5.2 DF auf Punkten

5.3 DF auf Binärdaten

5.4 DF auf Sequenzen

5.5 DF auf allgemeinen Mengen

- absoluter Betrag der Differenz zweier reeller Zahlen

$$d_{abs} : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}_0^+, d_{abs}(r_1, r_2) \mapsto |r_1 - r_2|$$

- euklidische Distanzfunktion d_{L_2} auf Punkten p_i der Menge \mathbb{R}^n

$$d_{L_2} : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}_0^+, d_{L_2}(p_1, p_2) \mapsto \sqrt{\sum_{i=1}^n (p_1[i] - p_2[i])^2}$$

Beispiel einer Pseudo-Distanzfunktion

5.1 Eigenschaften

5.2 DF auf Punkten

5.3 DF auf Binärdaten

5.4 DF auf Sequenzen

5.5 DF auf allgemeinen Mengen

indiskrete Pseudo-Distanzfunktion, die jedem Elementepaar aus $O \times O$ den Wert 0 zuweist:

$$d_{indiskret} : O \times O \longrightarrow \mathbb{R}_0^+, d_{indiskret}(o_1, o_2) \mapsto 0$$

(Funktion ist praktisch sinnlos)

Beispiel einer Semi-Distanzfunktion

Semi-Distanzfunktion d_{semi} auf der Menge $\{a, b, c\}$:

d_{semi}	a	b	c
a	0	1	3
b	1	0	1
c	3	1	0

Die Dreiecksungleichung ist nicht garantiert:

$$\begin{aligned}d_{semi}(a, c) &\not\leq d_{semi}(a, b) + d_{semi}(b, c) \\ 3 &\not\leq 1 + 1\end{aligned}$$

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Weitere Eigenschaften von Distanzfunktionen

5.1 Eigenschaften

5.2 DF auf Punkten

5.3 DF auf Binärdaten

5.4 DF auf Sequenzen

5.5 DF auf allgemeinen Mengen

folgende Eigenschaften werden an konkreten Funktionen getestet

- **Invarianz** bzgl.
 - Translation anhand Translationsobjekt T :
 $\forall o_1, o_2 : d(o_1, o_2) = d(o_1 + T, o_2 + T)$
 - Skalierung anhand Skalar S : $\forall o_1, o_2 : d(o_1, o_2) = d(o_1, S * o_2)$
 - Rotation anhand Rotationsobjekt R :
 $\forall o_1, o_2 : d(o_1, o_2) = d(R * o_1, R * o_2)$
- Darstellung des **Einheitskreises**: alle Punkte $o \in O$, für die $d(z, o) = 1$ gilt (z ist Zentrum)

Distanzeigenschaften im Einheitskreis

verschiedene Eigenschaften sind graphisch aus Einheitskreis erkennbar

- **Selbstidentität:** Zentrum liegt auf Kreis mit Radius 0.
- **Positivität:** Alle Punkte ungleich Zentrum liegen außerhalb des Kreises mit dem Radius 0
- **Translationsinvarianz:** Einheitskreis ändert Form nicht, wenn Zentrum verschoben wird
- **Symmetrie:** bei Translationsinvarianz und Symmetrie teilt Zentrum jede Diagonale zwischen zwei Randpunkten in genau zwei gleich lange Teile
- **Rotationsinvarianz:** Einheitskreis ist bezüglich Zentrum rotationssymmetrisch

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Überblick

5.1
Eigenschaften

5.1 Eigenschaften und Klassifikation

5.2 DF auf
Punkten

5.2 Distanzfunktionen auf Punkten

5.3 DF auf
Binärdaten

5.3 Distanzfunktionen auf Binärdaten

5.4 DF auf
Sequenzen

5.4 Distanzfunktionen auf Sequenzen

5.5 DF auf
allgemeinen
Mengen

5.5 Distanzfunktionen auf allgemeinen Mengen

Distanzfunktionen auf Punkten

Datentyp: array [1..n](real)

- Minkowski-Distanzfunktion L_m
- Gewichtete Minkowski-Distanzfunktion L_m^w
- Quadratische Distanzfunktion d_q
- Quadratische Pseudo-Distanzfunktion
- Dynamical-Partial-Semi-Pseudo-Distanzfunktion
- Chi-Quadrat-Semi-Pseudo-Distanzfunktion
- Kullback-Leibler-Abstandsfunktion
- Bhattacharyya-Abstandsfunktion

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Minkowski-Distanzfunktion L_m

am häufigsten eingesetzte Distanzfunktion auf Punkten mit $m > 0$

$$d_{L_m} : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}_0^+, d_{L_m}(p_1, p_2) \mapsto \left(\sum_{i=1}^n |p_1[i] - p_2[i]|^m \right)^{1/m}$$

$m = 1$: **Manhattan**-Distanzfunktion oder **Block**distanzfunktion

$m = 2$: **euklidische** Distanzfunktion

$m = \infty$: **Max**-Distanzfunktion oder **Tschebyscheff**-Distanzfunktion

Sonderfall bei $m = \infty$:

$$d_{L_\infty} = d_{L_{\max}}(p_1, p_2) \mapsto \max_{i=1}^n |p_1[i] - p_2[i]|$$

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Minkowski-Distanzfunktion L_m

Translationsinvarianz

T sei ein n -dimensionaler Vektor, der durch die Differenzberechnung aus der Formel verschwindet:

$$\begin{aligned}d_{L_m}(p_1 + T, p_2 + T) &= \left(\sum_{i=1}^n |(p_1[i] + T) - (p_2[i] + T)|^m \right)^{1/m} \\ &= \left(\sum_{i=1}^n |p_1[i] - p_2[i]|^m \right)^{1/m} \\ &= d_{L_m}(p_1, p_2)\end{aligned}$$

aber keine Skalierungs- oder Rotationsinvarianz

5.1
Eigenschaften

5.2 DF auf
Punkten

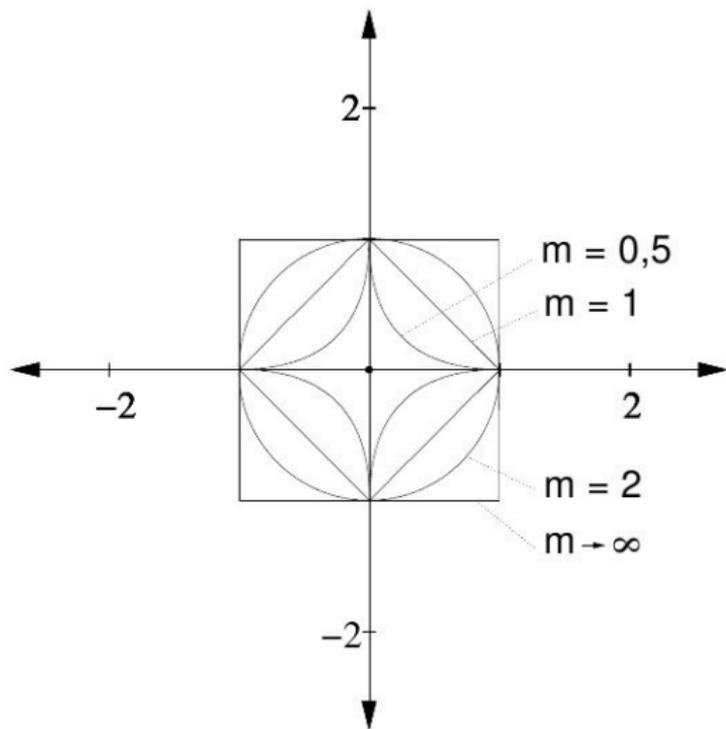
5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Minkowski-Distanzfunktion L_m

Einheitskreise



5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Minkowski-Distanzfunktion L_m

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Holdersche Ungleichung

es gilt immer:

$$(|a_1|^{m_1} + \dots + |a_n|^{m_1})^{1/m_1} \leq (|a_1|^{m_2} + \dots + |a_n|^{m_2})^{1/m_2} \text{ für } m_1 \geq m_2 \geq 1$$

also: Einheitskreis mit niedrigerem m -Wert liegt **innerhalb**
Einheitskreis mit höherem m -Wert

Minkowski-Distanzfunktion L_m

Sonderfall euklidische Distanzfunktion ($m = 2$)

- Einheitskreis ist **kreisförmig**
- **Rotationsinvarianz** ist erfüllt, da R Orthonormalmatrix ($R^T * R = R * R^T = I$)

$$\begin{aligned}d_{L_2}(R * p_1, R * p_2) &= \sqrt{(R * p_1 - R * p_2)^T * (R * p_1 - R * p_2)} \\ &= \sqrt{(p_1 - p_2)^T * R^T * R * (p_1 - p_2)} \\ &= \sqrt{(p_1 - p_2)^T * (p_1 - p_2)} \\ &= d_{L_2}(p_1, p_2)\end{aligned}$$

- Matrizeschreibweise: $d_{L_2}(p_1, p_2) = \sqrt{(p_1 - p_2)^T * (p_1 - p_2)}$

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Minkowski-Distanzfunktion L_m

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Berechnung von Reihenfolgen anhand Minkowski-Distanzfunktion
 L_m

Achtung: versch. m -Werte erzeugen untersch. Reihenfolgen!

$$p_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \text{ und } p_2 = \begin{pmatrix} 0,8 \\ 0,8 \end{pmatrix}.$$

Abstände dieser Punkte vom Koordinatenursprung O :

$$d_{L_1}(O, p_1) = 1 \text{ und } d_{L_1}(O, p_2) = 1.6$$

$$d_{L_\infty}(O, p_1) = 1 \text{ und } d_{L_\infty}(O, p_2) = 0,8$$

Gewichtete Minkowski-Distanzfunktion L_m^w

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

achsenparallele Stauchung und Streckung durch Gewichte $w_i \geq 0$

$$d_{L_m}^w : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}_0^+, d_{L_m}(p_1, p_2) \mapsto \left(\sum_{i=1}^n w_i * |p_1[i] - p_2[i]|^m \right)^{1/m}$$

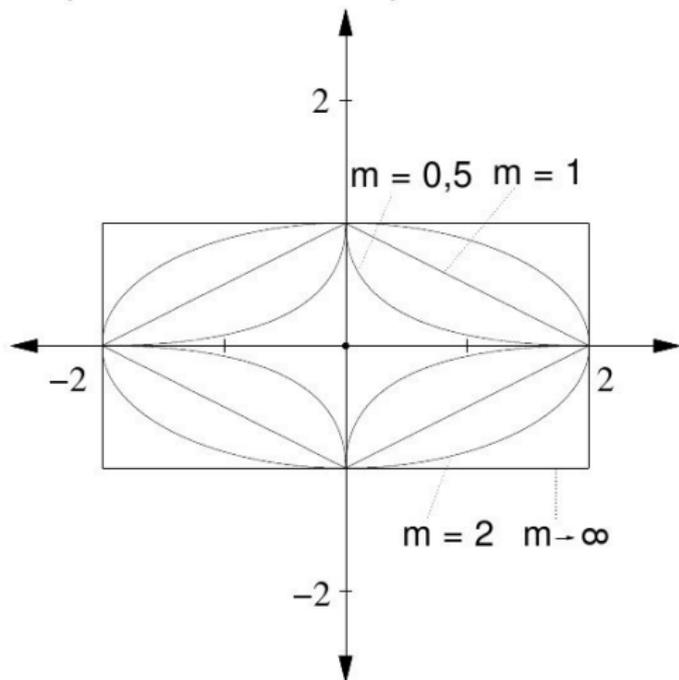
Forderung:

$$\sum_{i=1}^n w_i = 1.$$

Gewichtete Minkowski-Distanzfunktion L_m^w

Einheitskreis

$w_1 = 0.5, w_2 = 1$ (nicht normalisiert)



5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Invarianzen

- Translationsinvarianz
- keine Skalierungsinvarianz
- keine Rotationsinvarianz

Quadratische Distanzfunktion d_q

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Matrizenschreibweise:

$$d_q(p_1, p_2) = (p_1 - p_2)^T * A * (p_1 - p_2)$$

A im n -dimensionalen Raum ist **symmetrische** und **positiv** definite
Matrix $\mathbb{R}^{n \times n}$

Quadratische Distanzfunktion d_q

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Matrix A

- **Einheitsmatrix** E : d_q identisch mit $d_{L_2}^2$
- **Diagonalmatrix**: d_q entspricht $d_{L_2^w}^2$
(Gewichte korrespondieren zu Diagonalelementen)
- ansonsten: nonuniforme Skalierung, Rotation, Spiegelung der Punkte

Quadratische Distanzfunktion d_q

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Symmetrische und positiv definite Matrix A

es gilt immer: $A = U * L * U^T$ (Eigenwertzerlegung):

- U ist orthonormale Matrix (Rotation anhand Eigenvektoren)
- L ist Diagonalmatrix (Skalierung anhand Eigenwerten)

Quadratische Distanzfunktion d_q

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Symmetrische und positiv definite Matrix A

Berechnung der Distanz mittels $d_{L_2}^2$ auf transformierten Punkten
oft relativ schnell realisierbar

$$\begin{aligned}d_q(p_1, p_2) &= (p_1 - p_2)^T A (p_1 - p_2) \\&= (p_1 - p_2)^T U L U^T (p_1 - p_2) \\&= \left(L^{1/2} U^T (p_1 - p_2) \right)^T \left(L^{1/2} U^T (p_1 - p_2) \right) \\&= \left(L^{1/2} U^T p_1 - L^{1/2} U^T p_2 \right)^T \left(L^{1/2} U^T p_1 - L^{1/2} U^T p_2 \right) \\&= d_{L_2}^2 (L^{1/2} U^T p_1, L^{1/2} U^T p_2)\end{aligned}$$

Quadratische Distanzfunktion d_q

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Invarianzen

- Translationsinvarianz
- keine Skalierungsinvarianz
- keine Rotationsinvarianz

Quadratische Distanzfunktion d_q

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

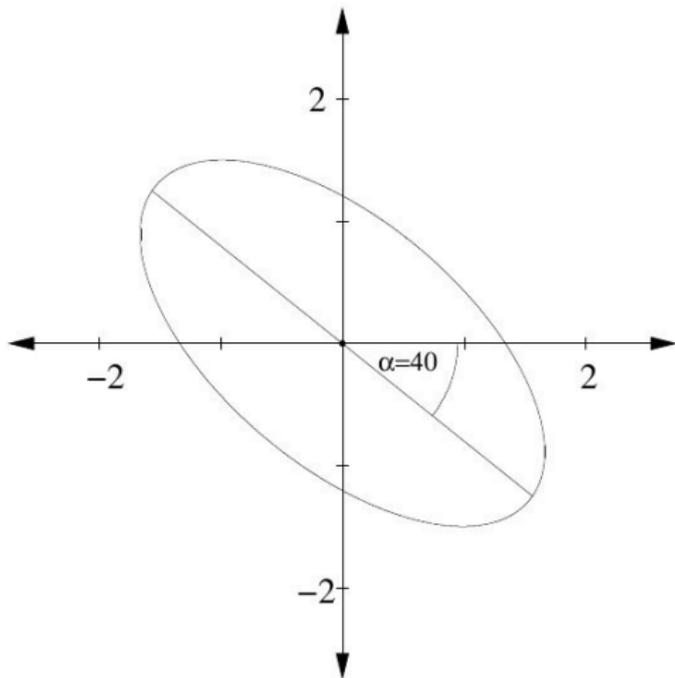
5.5 DF auf
allgemeinen
Mengen

Beispielmatrix

$$\begin{aligned} A &= \begin{pmatrix} 0,5599 & 0,3693 \\ 0,3693 & 0,6901 \end{pmatrix} \\ &= \begin{pmatrix} \cos 40 & \sin 40 \\ -\sin 40 & \cos 40 \end{pmatrix} * \begin{pmatrix} 0,25 & 0 \\ 0 & 1 \end{pmatrix} * \begin{pmatrix} \cos 40 & -\sin 40 \\ \sin 40 & \cos 40 \end{pmatrix} \end{aligned}$$

Quadratische Distanzfunktion d_q

Einheitskreis des Beispiels



5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Quadratische Distanzfunktion d_q

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Mahalanobis-Distanzfunktion

- Einsatz der quadratischen Distanzfunktion d_q , wenn Distanzberechnung Kombination unterschiedlicher Dimensionen erfordert
- Grundlage kann Kovarianzmatrix C auf d Dimensionen sein
→ **Mahalanobis**-Distanzfunktion $d_M(p_1, p_2)$

$$d_M(p_1, p_2) = |\det C|^{1/d} (p_1 - p_2)^T * C^{-1} * (p_1 - p_2)$$

Quadratische Pseudo-Distanzfunktion

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

- Aufgabe der Forderung nach Positiv-Definitheit für A
- Ziel: **unsymmetrische Translationsinvarianz** bzgl. Vektoren t des Vektorunterraums T :

$$pd_q(p_1, p_2 + t) = pd_q(p_1, p_2)$$

- Konstruktion der Matrix A aus geeigneter Orthogonalmatrix U und Diagonalmatrix L
- den U -Vektoren entsprechende Diagonalwerte auf Null setzen
- seien s_i mit $i = 1, \dots, m$ die durch l_i auf Null gesetzten U -Spaltenvektoren, dann gilt:

$$T = \left\{ t \in \mathbb{R}^n \mid t = \sum_{i=1}^m \lambda_i * s_i : \lambda_i \in \mathbb{R} \right\}$$

Quadratische Pseudo-Distanzfunktion

Nachweis der Translationsinvarianz

$$\begin{aligned} & pd_q(p_1, p_2 + t) \\ = & (p_1 - p_2 - t)^T U L U^T (p_1 - p_2 - t) \\ = & (p_1 - p_2 - t)^T U L^{1/2} L^{1/2} U^T (p_1 - p_2 - t) \\ = & \left(L^{1/2} U^T (p_1 - p_2 - t) \right)^T \left(L^{1/2} U^T (p_1 - p_2 - t) \right) \\ = & \left(L^{1/2} U^T (p_1 - p_2) - L^{1/2} U^T t \right)^T \left(L^{1/2} U^T p_1 - p_2) - L^{1/2} U^T t \right) \\ = & \left(L^{1/2} U^T (p_1 - p_2) \right)^T \left(L^{1/2} U^T p_1 - p_2) \right) \\ = & pd_q(p_1, p_2) \end{aligned}$$

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Quadratische Pseudo-Distanzfunktion

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Beispiel Quadratische Pseudo-Distanzfunktion

Konstruktion Translationsinvarianz im Winkel 40 Grad:

$$U = \begin{pmatrix} \cos 40 & -\sin 40 \\ \sin 40 & \cos 40 \end{pmatrix}$$

$$L = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

Die Kombination dieser Matrizen ergibt die gewünschte Matrix A :

$$U * L * U^T = \begin{pmatrix} 0,4132 & -0,4924 \\ -0,4924 & 0,5868 \end{pmatrix}$$

Quadratische Pseudo-Distanzfunktion

Einheitskreis des Beispiels

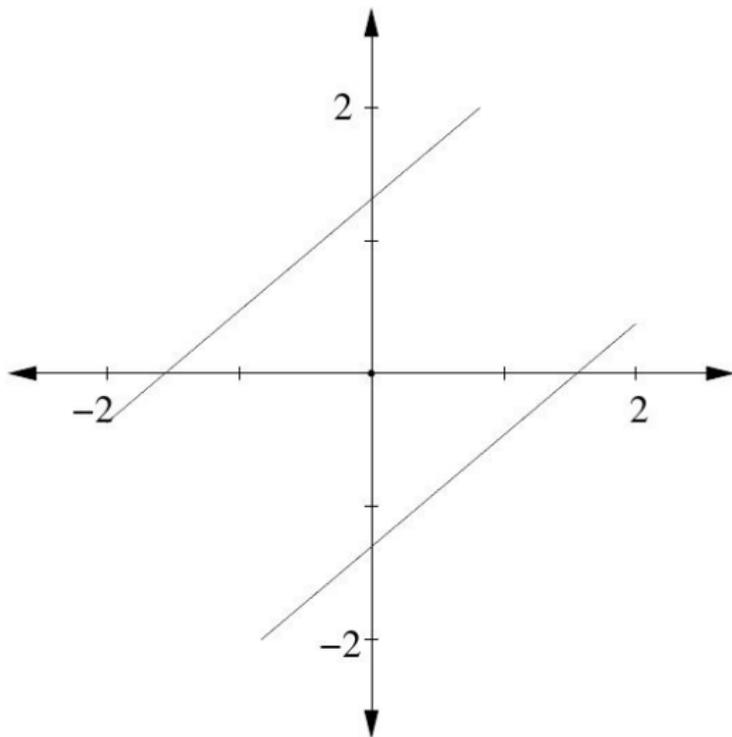
5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen



Dynamical-Partial-Semi-Pseudo-Distanzfunktion

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

folgende Beobachtungen Chang/Wu03 bzgl. Unähnlichkeit im hochdimensionalen Raum:

- ähnliche Objekte liegen meist nur in wenigen Dimensionen nebeneinander
- Ähnlichkeit kann häufig nicht an bestimmten Dimensionen festgemacht werden

Problem mit Minkowski-Distanzfunktion: **alle** Dimensionen werden berücksichtigt

Dynamical-Partial-Semi-Pseudo-Distanzfunktion

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

- Berücksichtigung einer **dynamischen** Untermenge der Dimensionen
- p_1 und p_2 seien zwei Punkte im n -dimensionalen Raum und $\delta_i = |p_1[i] - p_2[i]|$ der Abstand in Dimension i
- nur die **m kleinsten** Abstände werden berücksichtigt:
 $\Delta_m = \{\text{die kleinsten } m \text{ } \delta\text{-Werte aus } (\delta_1, \delta_2, \dots, \delta_n)\}$

$$d_{dp}^{m,r} = \left(\sum_{\delta_i \in \Delta_m} \delta_i^r \right)^{\frac{1}{r}}$$

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

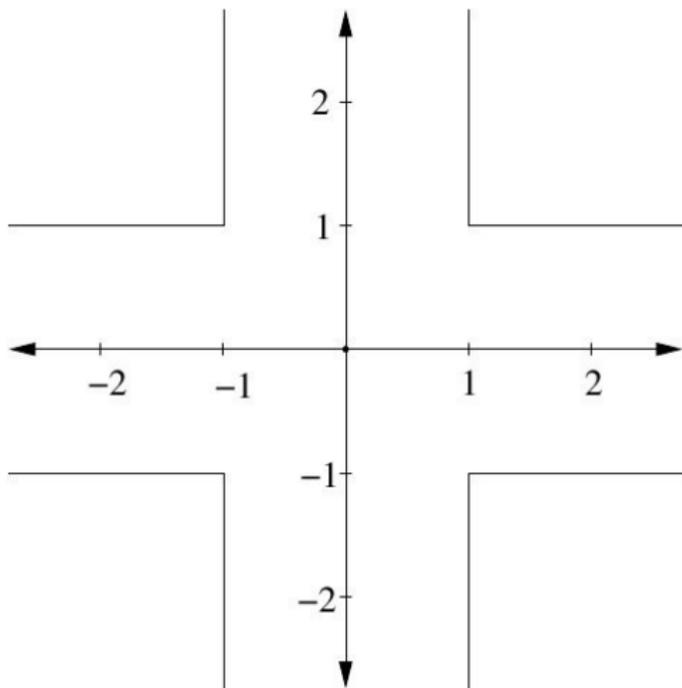
Eigenschaften

- Selbstidentität und Symmetrie sind erfüllt
- Verletzung der Positivität und Dreiecksungleichung

Dynamical-Partial-Semi-Pseudo-Distanzfunktion

Einheitskreis

zweidimensionaler Raum und $m = 1$



5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Chi-Quadrat-Semi-Pseudo-Distanzfunktion

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

- Abstand zwischen **Histogrammen mit absoluten Häufigkeiten**
- ursprünglich in Statistik entwickelt Untersuchung von Abhängigkeiten zwischen Zufallsvariablen
- basiert auf **Nullhypothese**: Häufigkeitsverteilungen sind gleich also Differenz zwischen erwarteter und tatsächlicher Häufigkeiten sind 0

Chi-Quadrat-Semi-Pseudo-Distanzfunktion

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

$$spd_{\chi^2}(p_1, p_2) = \sum_{j=1}^n \frac{(p_1[j] - \hat{p}_1[j])^2}{\hat{p}_1[j]} + \sum_{j=1}^n \frac{(p_2[j] - \hat{p}_2[j])^2}{\hat{p}_2[j]} \text{ für } p_1, p_2 \in \mathbb{N}_0^n$$

erwartete Häufigkeiten:

$$\hat{p}_i[j] = \frac{(p_1[j] + p_2[j]) * \sum_{a=1}^n p_i[a]}{\sum_{a=1}^n (p_1[a] + p_2[a])}.$$

Chi-Quadrat-Semi-Pseudo-Distanzfunktion

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Beispiel

- Test, ob Grippedoppelimpfung Grippe verhindern kann
- Befragung verschiedener Personen über Auftreten von Grippe und Impfungen
- erwartete Werte sind in Klammern notiert

	keine Impfung	eine Impfung	Doppelimpfung	Σ
Grippe	24 (14,398)	9 (5,014)	13 (26,588)	46
keine Grippe	289 (298,602)	100 (103,986)	565 (551,412)	954
Σ	313	109	578	1000

Chi-Quadrat-Semi-Pseudo-Distanzfunktion

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Berechnung der erwarteten Häufigkeiten

- wenn kein Zusammenhang zwischen Impfung und Grippe, dann Wert jeder Zelle abschätzbar
- Beispiel Grippe/keine Impfung:
 - Wahrscheinlichkeit für Grippe ist $46/1000$
 - Wahrscheinlichkeit für keine Impfung ist $313/1000$
 - Wahrscheinlichkeit für Grippe/keine Impfung:
 $46/1000 * 313/1000$
 - erwartete Häufigkeit:
 $46/1000 * 313/1000 * 1000 = 46 * 313 / 1000 = 14,398$

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

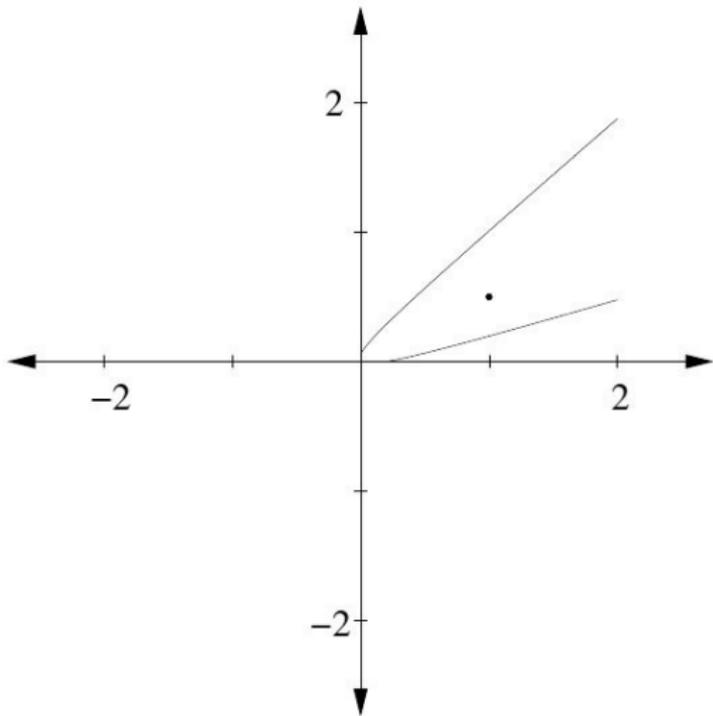
5.5 DF auf
allgemeinen
Mengen

Eigenschaften

- Selbstidentität und Symmetrie sind erfüllt
- Rotationsinvarianz
- keine Positivität
- keine Dreiecksungleichung

Chi-Quadrat-Semi-Pseudo-Distanzfunktion

Einheitskreis



5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Überblick

5.1
Eigenschaften

5.1 Eigenschaften und Klassifikation

5.2 DF auf
Punkten

5.2 Distanzfunktionen auf Punkten

**5.3 DF auf
Binärdaten**

5.3 Distanzfunktionen auf Binärdaten

5.4 DF auf
Sequenzen

5.4 Distanzfunktionen auf Sequenzen

5.5 DF auf
allgemeinen
Mengen

5.5 Distanzfunktionen auf allgemeinen Mengen

Allgemeines

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

- ▶ Unter Binärdaten verstehen wir hier die Erfüllung bzw. Nichterfüllung bestimmter Eigenschaften von Medienobjekten.
- ▶ Zu einer vorgegebenen Menge E von n Eigenschaften und jedem Medienobjekt bekannt ist, welche Eigenschaften erfüllt sind und welche nicht.
- ▶ Graphisch lassen sich die Feature-Daten als Eckpunkte eines n -dimensionalen Hypereinheitswürfels darstellen.

Eigenschaften und Korrespondenzen

Vergleicht man zwei Punkte p_1 und p_2 , ergeben sich vier verschiedene Anzahlwerte, die als Grundlage für die Distanzmessung verwendet werden:

$e \in E$	e erfüllt für p_1	e nicht erfüllt für p_1
e erfüllt für p_2	$n_{1/1}$	$n_{0/1}$
e nicht erfüllt für p_2	$n_{1/0}$	$n_{0/0}$

Beispiel:

$$p_1 = (0, 0, 0, 0, 1, 1, 1, 1)^T \quad p_2 = (1, 1, 0, 1, 1, 1, 0, 0)^T$$

↓

$$n_{0/0} = 1 \quad n_{0/1} = 3 \quad n_{1/0} = 2 \quad n_{1/1} = 2$$

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Minkowski-Distanzfunktion auf Binärdaten

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

In der allgemeinen Form:

$$d_{L_m}(p_1, p_2) = \left(\sum_{i=1}^n |p_1[i] - p_2[i]|^m \right)^{1/m}$$

Auf Binärdaten:

$$d_{L_m} = (n_{1/0} + n_{0/1})^{1/m}$$

Überblick

5.1
Eigenschaften

5.1 Eigenschaften und Klassifikation

5.2 DF auf
Punkten

5.2 Distanzfunktionen auf Punkten

5.3 DF auf
Binärdaten

5.3 Distanzfunktionen auf Binärdaten

**5.4 DF auf
Sequenzen**

5.4 Distanzfunktionen auf Sequenzen

5.5 DF auf
allgemeinen
Mengen

5.5 Distanzfunktionen auf allgemeinen Mengen

Allgemeines

- ▶ Sequenz-Daten bestehen aus einer Liste von Datenelementen eines Datentyps.
- ▶ Die Anzahl der Elemente zur Beschreibung von verschiedenen Medienobjekten kann unterschiedlich sein.
- ▶ Klassifikation und Beispiele:
 - ▶ keine Positionskorrespondenz:
Earth-Mover-Distanzfunktion
 - ▶ Positionskorrespondenz und reelle Werte:
DFT- L_2 -Distanzfunktion
 - ▶ Positionskorrespondenz und nominale Werte:
Editierdistanzfunktion

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Earth-Mover-Distanzfunktion

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

- ▶ Diese Distanzfunktion geht von dem Datentyp $\text{tuple}(p_i : \text{array}[1 \dots n](\text{real}) ; w_{p_i} : \text{real})$ für das i -te Element einer Sequenz p aus.

- ▶ “Erdhügel und Erdlöcher”

Um die Distanz zwischen der Sequenz p mit m Elementen und der Sequenz q mit n Elementen zu ermitteln, werden die Elemente von p als Erdhügel und die von q als Erdlöcher aufgefasst. Die Punkte p_i bzw. q_i geben die Position der Hügel bzw. der Löcher an, während w_{p_i} und w_{q_i} die Volumina der Hügel/Löcher beschreiben. Um die Distanz zwischen den Sequenzen zu ermitteln, wird nun versucht, die Erde der Hügel mit minimalen Transportkosten in die Löcher zu füllen.

Earth-Mover-Distanzfunktion

- ▶ Ziel ist Minimierung der Transportkosten.
- ▶ Ein konkreter Transport wird durch die Angabe der Quantitätsmatrix $F = [f_{ij}]$ definiert. Der Wert f_{ij} gibt die Menge der Erde an, die vom Hügel p_i zum Loch q_j transportiert wird.
- ▶ Die Transportkosten berechnen sich aus dem Produkt der Quantitäten mit den entsprechenden Grunddistanzwerten:

$$\text{Kosten}(p, q, F) = \sum_{i=1}^m \sum_{j=1}^n d(p_i, q_j) f_{ij}$$

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Earth-Mover-Distanzfunktion

Diese Kosten sind unter Berücksichtigung folgender Bedingungen zu minimieren:

$$f_{ij} \leq 0 \quad : \quad 1 \leq i \leq m, 1 \leq j \leq n \quad (1)$$

$$\sum_{j=1}^n f_{ij} \leq w_{p_i} \quad : \quad 1 \leq i \leq m \quad (2)$$

$$\sum_{i=1}^m f_{ij} \leq w_{q_j} \quad : \quad 1 \leq j \leq n \quad (3)$$

$$\sum_{i=1}^m \sum_{j=1}^n f_{ij} = \min \left(\sum_{i=1}^m w_{p_i}, \sum_{j=1}^n w_{q_j} \right) \quad (4)$$

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Earth-Mover-Distanzfunktion

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Der Distanzwert nach der EM-Distanzfunktion berechnet sich nun aus der Normierung der minimalen Transportkosten bezüglich der Gesamtmenge der transportierten Erde:

$$d_{EM}(p, q) = \frac{\min_{|f_{ij}|} \left(\sum_{i=1}^m \sum_{j=1}^n d(p_i, q_j) f_{ij} \right)}{\sum_{i=1}^m \sum_{j=1}^n f_{ij}}$$

DFT- L_2 -Distanzfunktion

- ▶ Diese Distanzfunktion ist zum Vergleich von Sequenzen reeller Werte mit einer Positionskorrespondenz und fester Länge geeignet.
- ▶ Die Grundidee liegt in der Verwendung der euklidischen Distanzfunktion auf den einzelnen korrespondierenden Sequenzwerten.
- ▶ Ein typisches Beispiel für Sequenzen, bei denen diese Distanzfunktion sinnvoll angewendet werden kann, sind Zeitreihen. Z. B. können Tierpopulationen, Pegelstände, aber auch Aktienkursverläufe als zeitabhängige Werte miteinander verglichen werden.

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

Editierdistanz

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

- ▶ Ein Beispiel für eine Distanz zwischen Sequenzen mit nominalen Werten anhand einer abgeschwächten Positionskorrespondenz.
- ▶ Die Editierdistanz misst den minimalen Aufwand, um eine Zeichenkette mittels Editieroperationen in eine andere Zeichenkette zu überführen.
- ▶ Beispiel - die Editierdistanz zwischen den Wörtern "Abend" und "Robe" beträgt 4 (Ersetzen von "A" durch "R", Einfügen von "o", Entfernen von "n" und "d").

Überblick

5.1
Eigenschaften

5.1 Eigenschaften und Klassifikation

5.2 DF auf
Punkten

5.2 Distanzfunktionen auf Punkten

5.3 DF auf
Binärdaten

5.3 Distanzfunktionen auf Binärdaten

5.4 DF auf
Sequenzen

5.4 Distanzfunktionen auf Sequenzen

**5.5 DF auf
allgemeinen
Mengen**

5.5 Distanzfunktionen auf allgemeinen Mengen

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen

- ▶ Bisher waren die Datentypen der zu vergleichenden Objekte eingeschränkt. Jetzt soll ein Vergleich auf allgemeinen Mengen erfolgen.
- ▶ Beispiele:
 - ▶ die Bottleneck-Distanzfunktion,
 - ▶ die Distanzfunktion über das Volumen der symmetrischen Differenz,
 - ▶ die Hausdorff-Distanzfunktion, und
 - ▶ die Frechet-Distanzfunktion

Bottleneck-Distanzfunktion

- ▶ d_B ist auf endlichen Untermengen einer Menge X mit einer gegebenen Grunddistanz $d_X : X \times X \rightarrow \mathbb{R}_0$ definiert.

- ▶ Die Kardinalität beider Untermengen muss gleich sein

$$A, B \subset X \quad \text{mit} \quad |A| = |B|$$

- ▶ Zwischen den Elementen der Untermengen existiere eine bijektive Abbildung f . Man ist an der Distanz d_X des am weitesten auseinanderliegenden Elementepaars interessiert. Das Minimum der maximalen Elementepaarabstände über allen möglichen Bijektionen $f \in F(A, B)$ wird gesucht:

$$d_B(A, B) = \min_{f \in F(A, B)} \max_{a \in A} d_X(a, f(a))$$

5.1
Eigenschaften

5.2 DF auf
Punkten

5.3 DF auf
Binärdaten

5.4 DF auf
Sequenzen

5.5 DF auf
allgemeinen
Mengen