

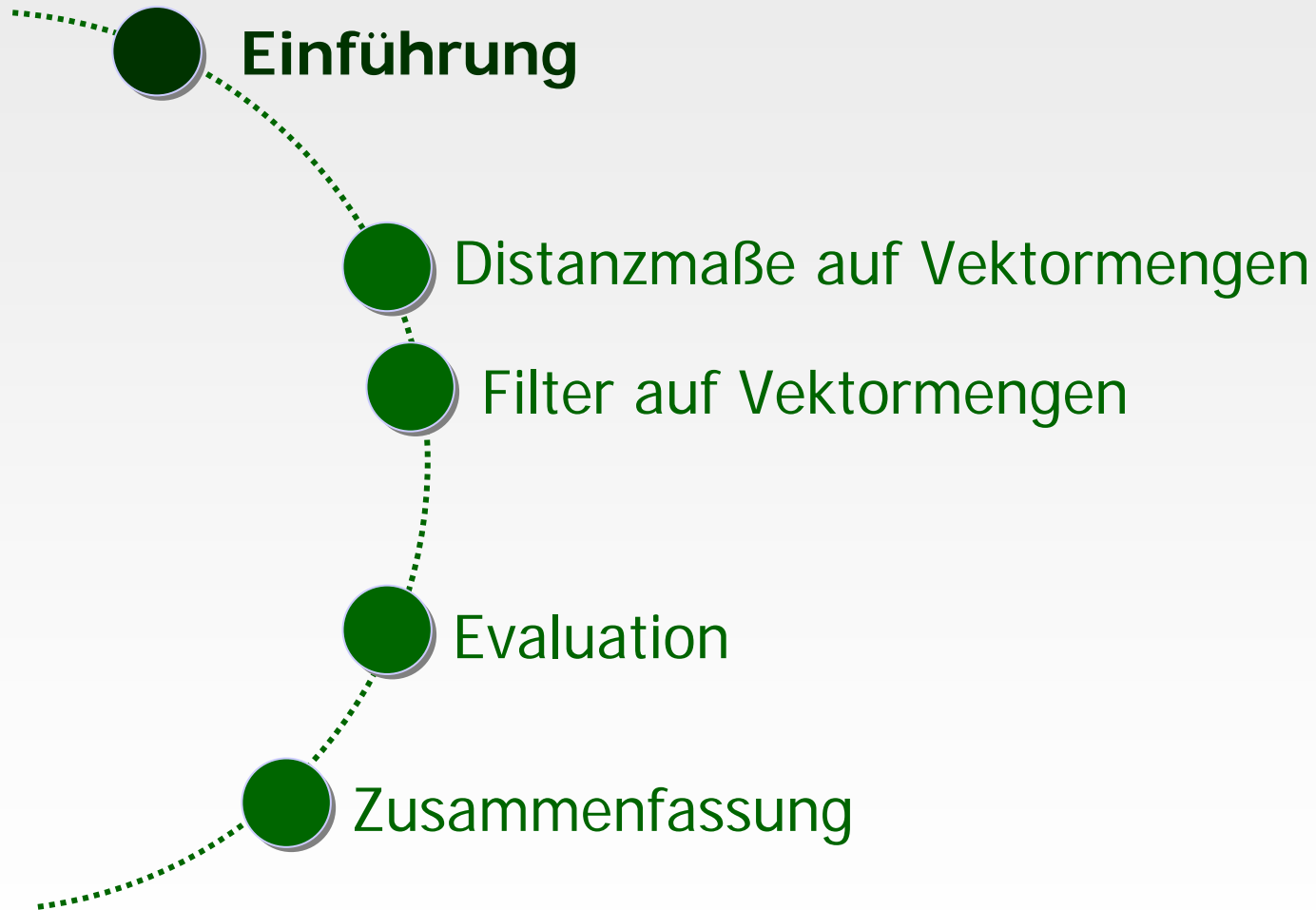


Efficient Similarity Search on Vector Sets

Stefan Brecheisen, Hans-Peter Kriegel, Martin Pfeifle

Lehr- und Forschungseinheit für Datenbanksysteme
Institut für Informatik
Ludwig-Maximilians-Universität München

Überblick



Einführung

Featurebasierte Ähnlichkeit:

- Extraktion von d -dimensionalen Featurevektoren
- Ähnlichkeitsmaß ist die Distanz im Feature-Raum

Distanzbasierte Ähnlichkeit:

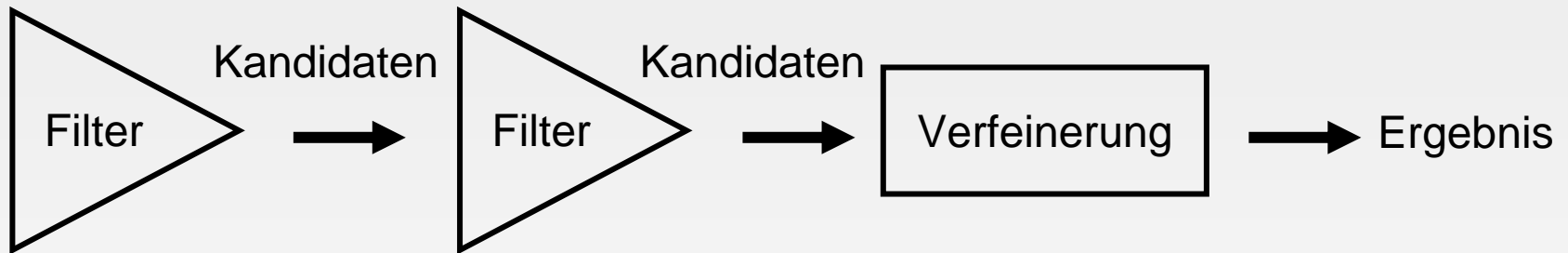
- Ähnlichkeitsmaß ist direkt auf den Objekten definiert, z.B. Bäume, Graphen

Mittelweg:

- Fasse pro Objekt k Featurevektoren zu einer Menge zusammen
- Ähnlichkeitsmaß ist die Distanz zwischen den Vektormengen

Einführung

Effizienzsteigernde Maßnahmen:

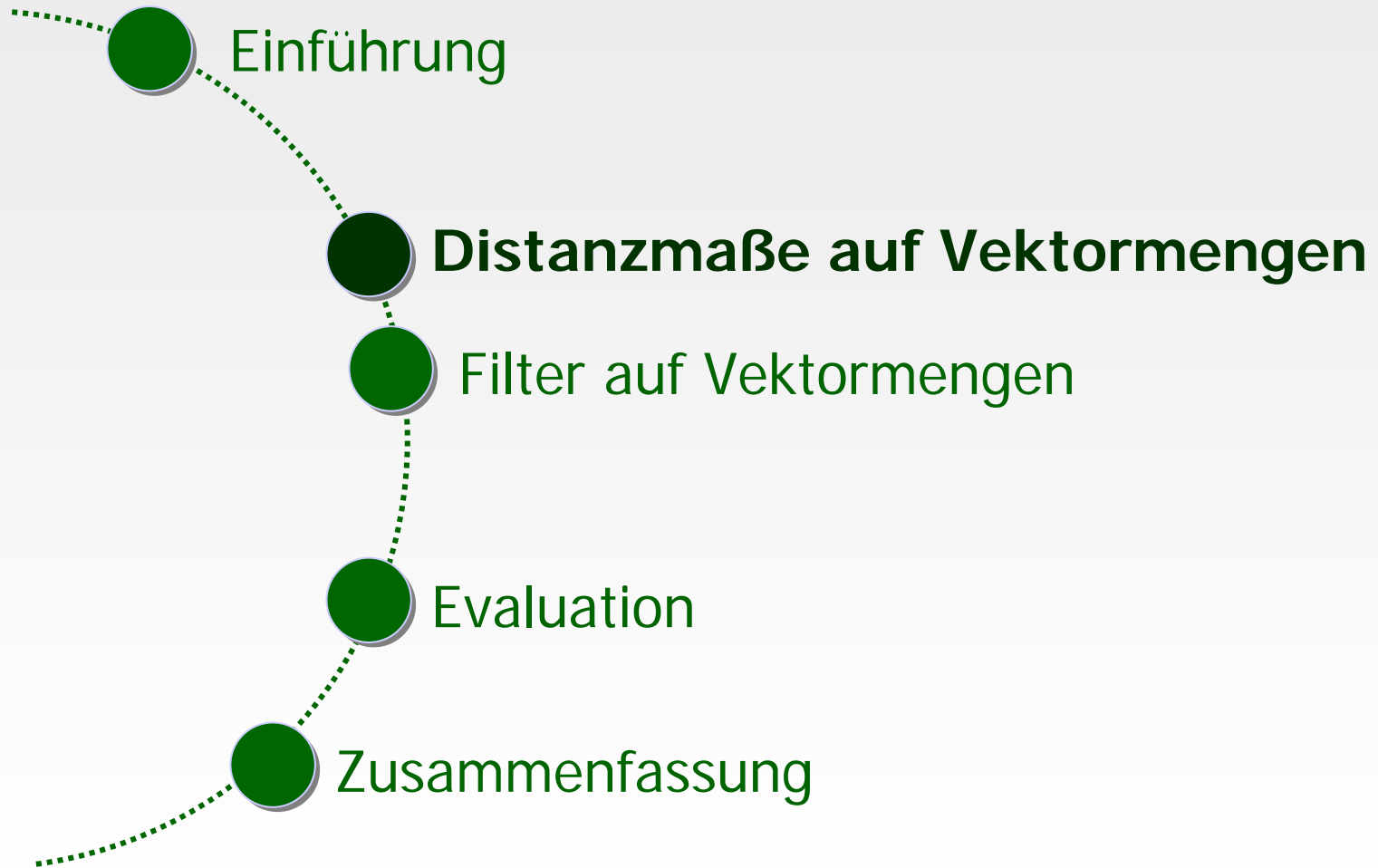


- Mehrstufige Anfragebearbeitung
→ Untere-Schranke-Eigenschaft

$$\forall o_1, o_2 \in O : d_f(o_1, o_2) \leq d_o(o_1, o_2)$$

- Indexstrukturen
→ Metrische Indexstrukturen, z.B. M-tree

Überblick



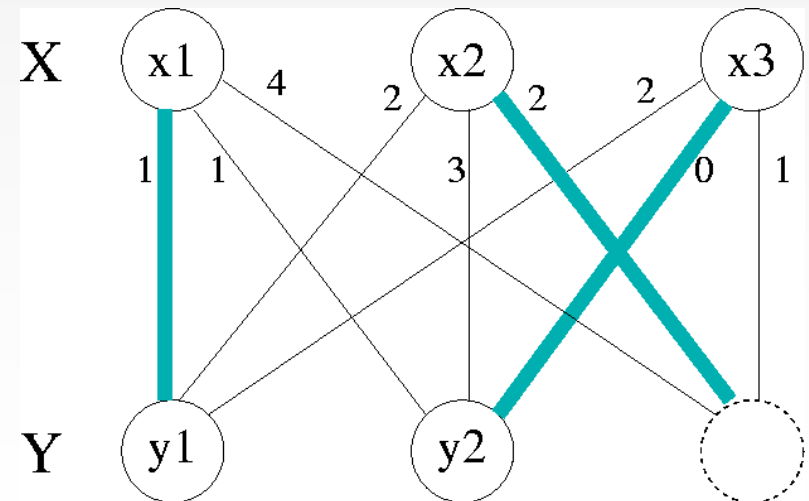
Distanzmaße auf Vektormengen

Anforderungen an Distanzmaß:

- Eignung für Ähnlichkeitssuche
- Metrik

→ Bestimme maximales Matching mit minimalem Gewicht
(Minimal-Matching-Distanz)

- Gegeben Vektormengen X und Y ,
 $|Y| \leq |X| \leq k$
- Konstruiere vollständigen bipartiten
Graph $G = (X \cup Y, X \times Y)$
- Das Gewicht jeder Kante
 $(x,y) \in X \times Y$ ist $dist(x,y)$
- Gewichtsfunktion w für nicht
zugeordnete Knoten, falls $|X| \neq |Y|$



Distanzmaße auf Vektormengen

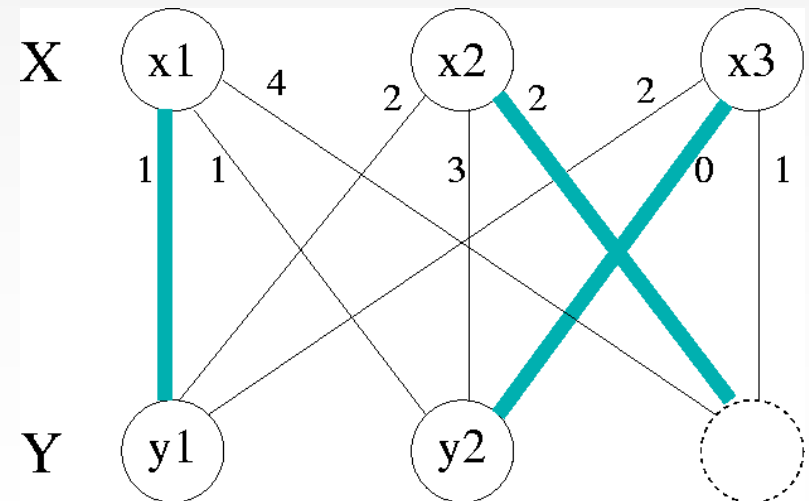
Partielle Minimal-Matching-Distanz:

- Bestimme Zuordnung zwischen $s \leq |X|$ Vektoren

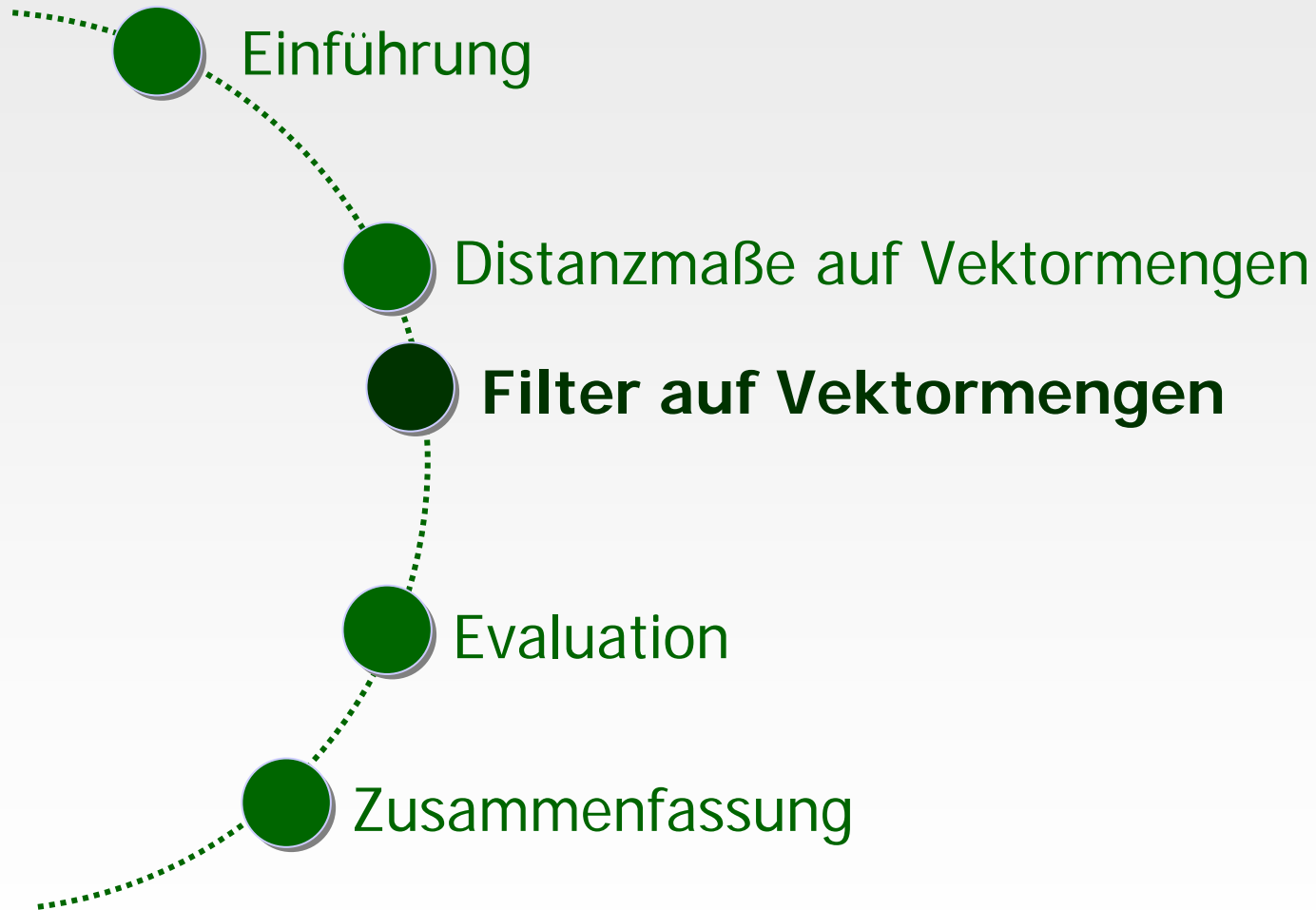
Zeitaufwand:

- für vollständiges Matching $O(k^3 + k^2d)$
- für partielles Matching $O(\binom{k}{s}sk^2 + k^2d)$

- Gegeben Vektormengen X und Y ,
 $|Y| \leq |X| \leq k$
- Konstruiere vollständigen bipartiten
Graph $G = (X \cup Y, X \times Y)$
- Das Gewicht jeder Kante
 $(x,y) \in X \times Y$ ist $dist(x,y)$
- Gewichtsfunktion w für nicht
zugeordnete Knoten, falls $|X| \neq |Y|$



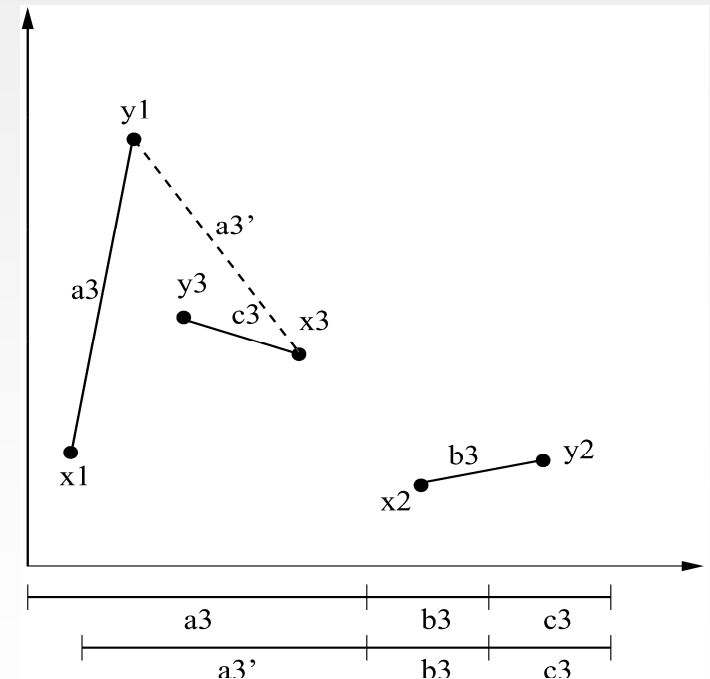
Überblick



Filter auf Vektormengen

Closest-Pair-Ansatz:

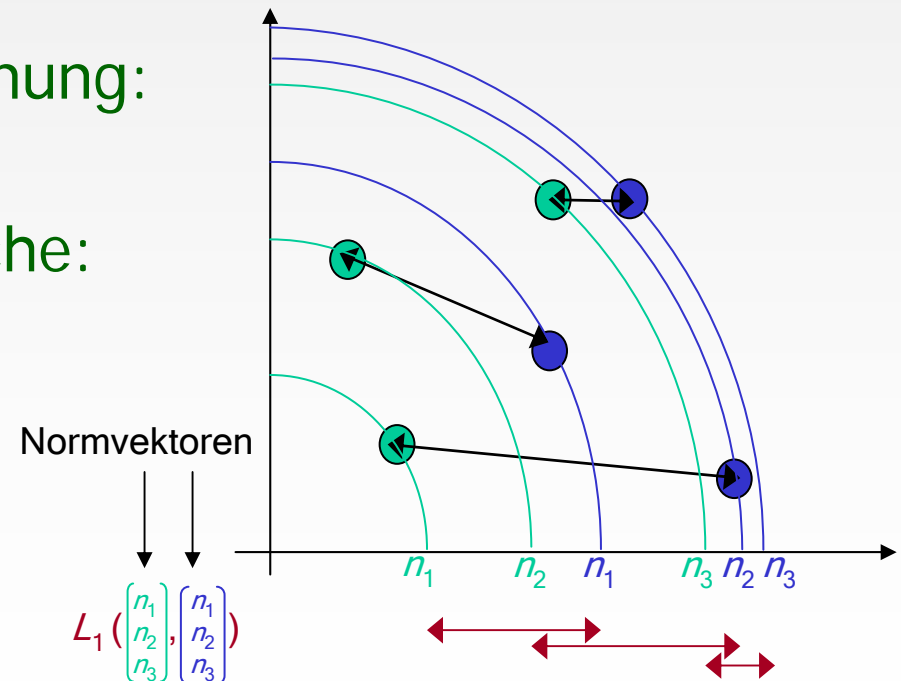
- Summieren der Abstände zu den nächsten Nachbarn in der anderen Menge
- 1:n-Zuordnungen sind zugelassen
- Kosten einer Distanzberechnung: $O(k^2d)$
- Für partielle Ähnlichkeitssuche: Summieren der s kürzesten Abstände zwischen den nächsten Nachbarn



Filter auf Vektormengen

Norm-Vektor-Ansatz:

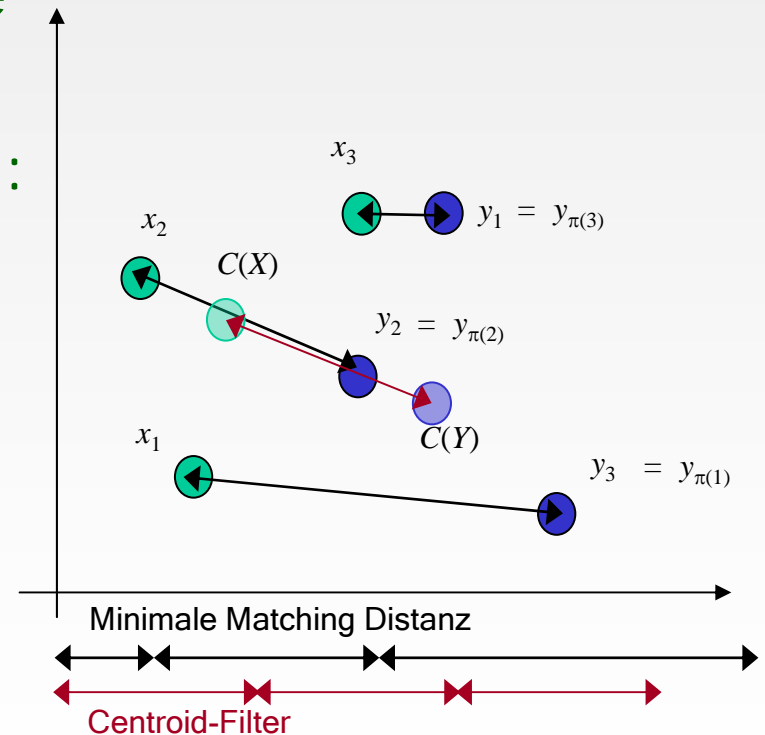
- Aggregation über die Dimensionen
- Zusammenfassen der sortierten Normwerte im Normvektor
- Kosten einer Distanzberechnung: $O(k)$
- Für partielle Ähnlichkeitssuche: Closest-Pair-Ansatz auf den Normwerten



Filter auf Vektormengen

Centroid-Ansatz:

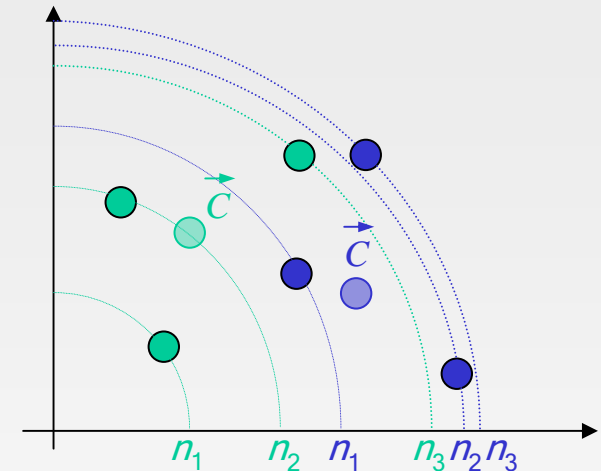
- Aggregation über die Vektoren
- Approximation der Vektormenge durch ihren Centroid
- Kosten einer Distanzberechnung: $O(d)$
- Für partielle Ähnlichkeitssuche nicht anwendbar



Filter auf Vektormengen

Kombinierter Ansatz:

- Aggregation über Punkte und Dimensionen
- Kosten einer Distanzberechnung: $O(d+k)$

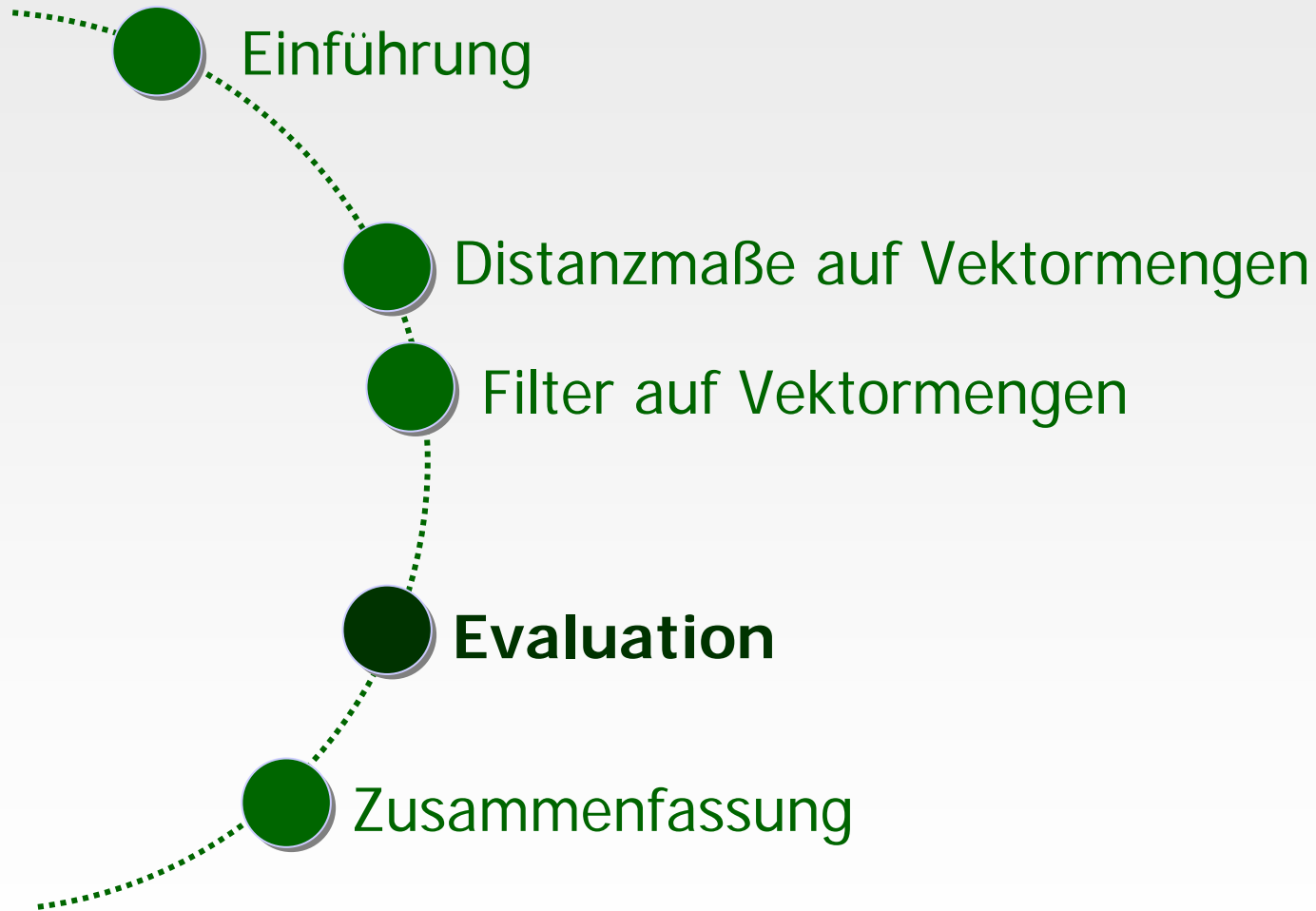


$$\max \left\{ L_1 \left(\begin{bmatrix} n_1 \\ n_2 \\ n_3 \end{bmatrix}, \begin{bmatrix} n_1 \\ n_2 \\ n_3 \end{bmatrix} \right), L_p \left(\vec{C}, \vec{C} \right) \right\}$$

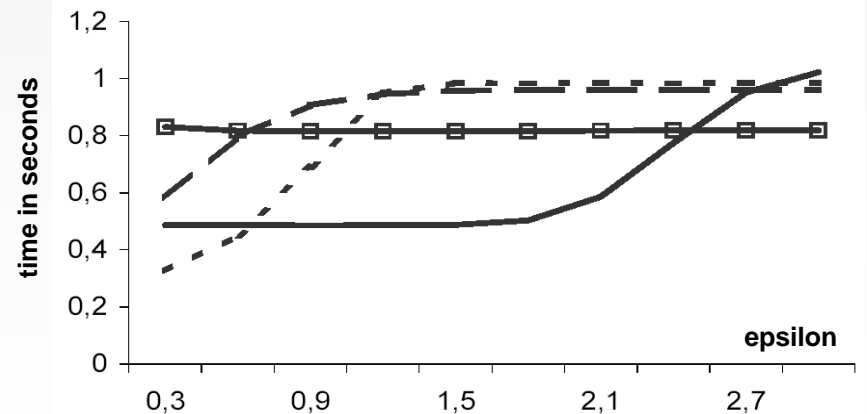
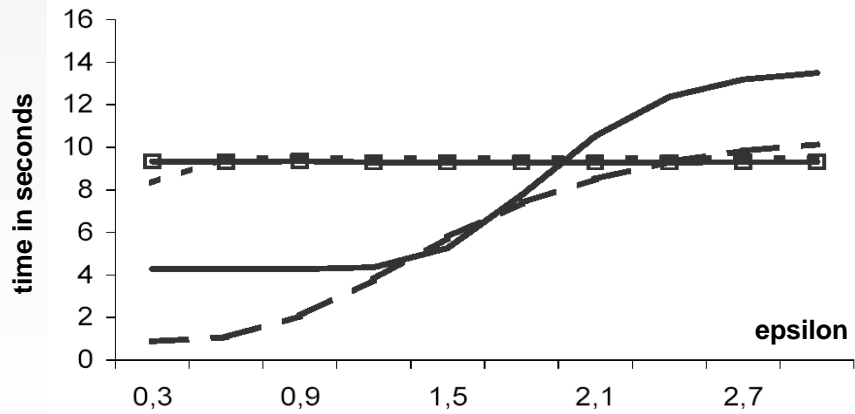
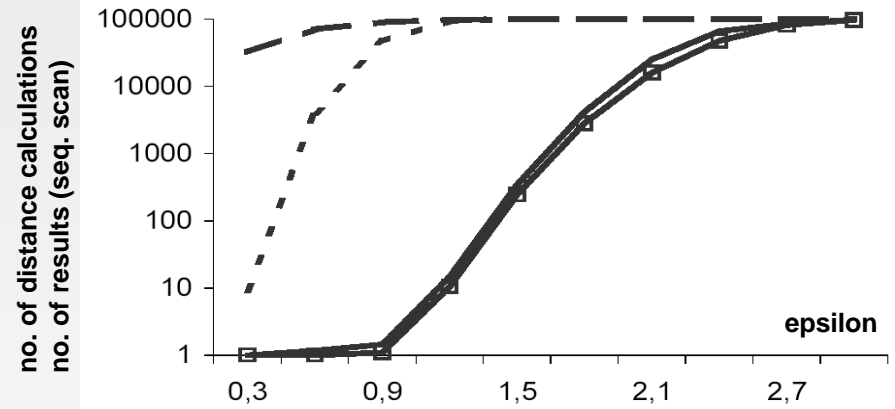
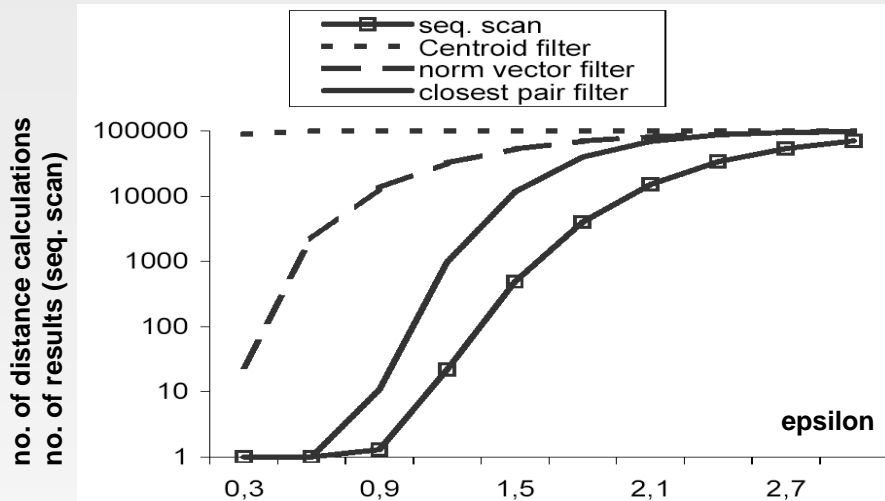
Zusammenfassung:

	exact distance	closest pair	centroid	norm vector
complete similarity	$O(k^3 + k^2 d)$	$O(k^2 d)$	$O(d)$	$O(k)$
partial similarity	$O(\binom{k}{s} s k^2 + k^2 d)$	$O(k^2 d \log s)$	n/a	$O(k \log s)$

Überblick



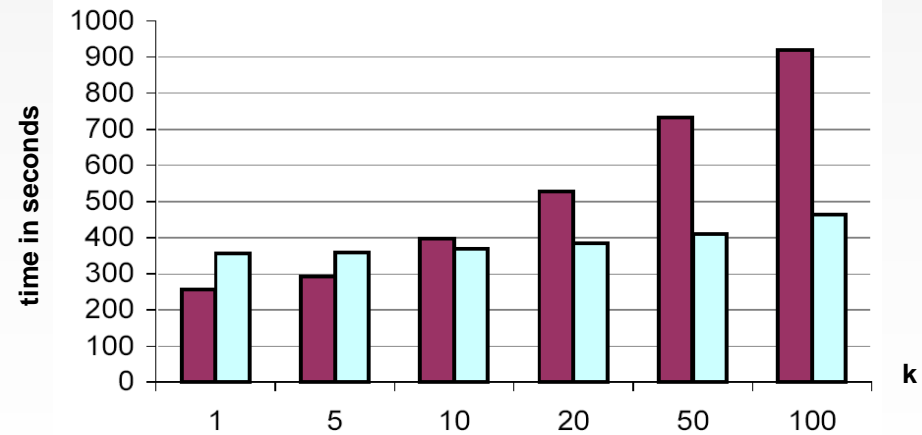
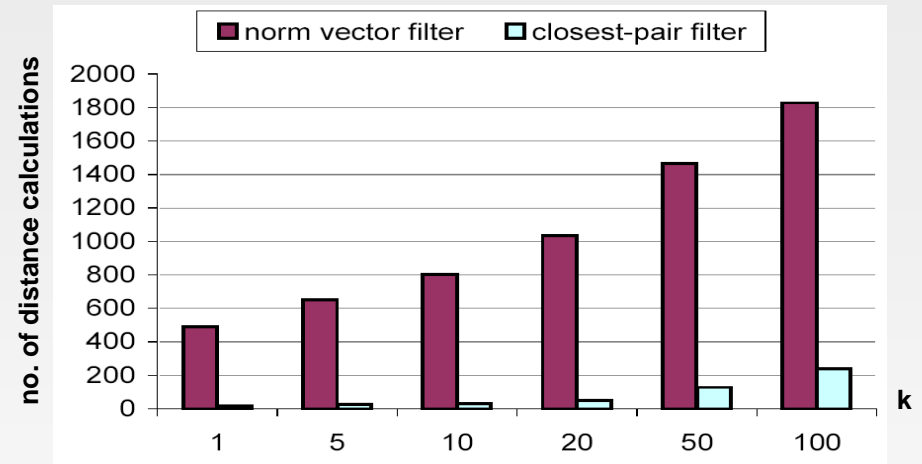
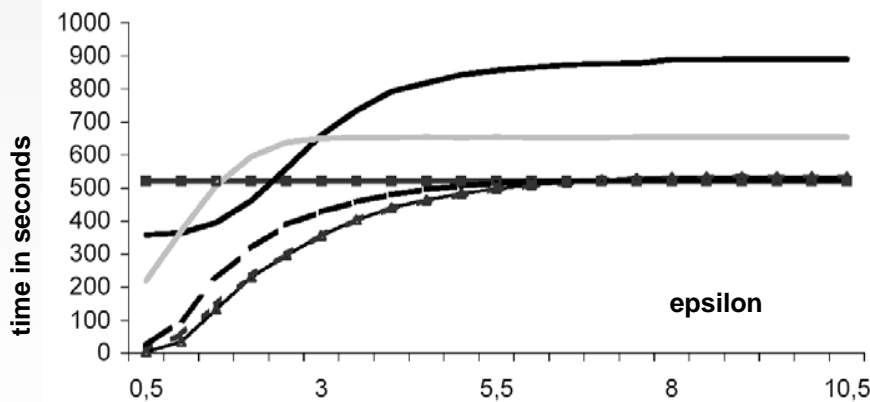
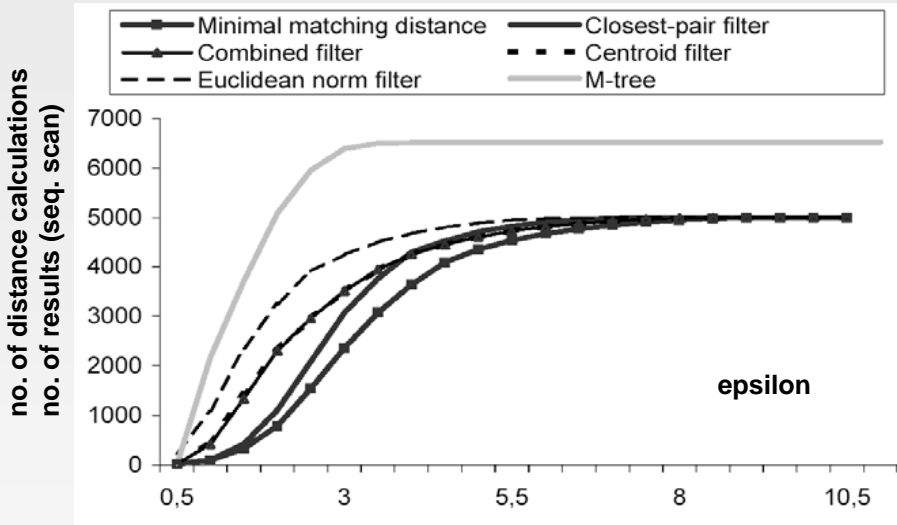
Evaluation – Künstliche Daten



10 Vektoren, 2 Dimensionen

2 Vektoren, 10 Dimensionen

Evaluation - Realdaten



vollständige Bereichsanfragen

partielle knn-Anfragen

Überblick



Zusammenfassung

- Datenmodellierung durch Vektormengen
- Distanzmaß auf Vektormengen für vollständige und partielle Ähnlichkeitssuche
- Mehrere Filter unterschiedlicher Komplexität zur Anfragebeschleunigung

Ausblick:

- Vektormengen und mehrstufige Anfragebearbeitung für Data Mining, z.B. Clustering und Klassifikation

?

?

?

?

?

?

?

?

?

Fragen?