Kapitel 2 Ähnlichkeitsmodelle

2.1 Allgemeine Konzepte

Verschiedene Ähnlichkeitsmodelle, um subjektive Ähnlichkeitsbegriffe zu objektivieren:

- Allgemeine Ähnlichkeitsmodelle
  - Beispiel: Ähnlichkeit ist Anteil übereinstimmender Eigenschaften zweier Objekte
  - Übereinstimmende Merkmale führen z.B. zu “100% Ähnlichkeit”

- Distanzbasierte Ähnlichkeit
  Der Wert einer Distanzfunktion beschreibt die (Un-)Ähnlichkeit von Objekten.
  - Je größer die Distanz, desto unähnlicher sind die Objekte.
  - Ein Objekt q hat zu sich selbst den Abstand Null, d.h. aus $p = q$ folgt $d(p, q) = 0$.

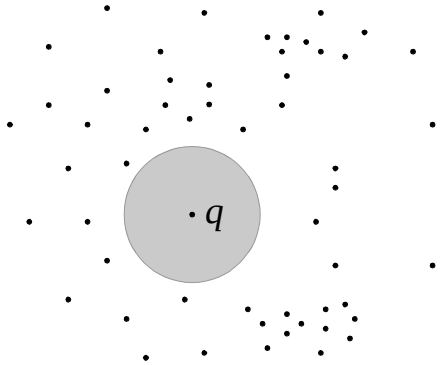
2.1.1 Typen von Ähnlichkeitsanfragen

Basis: Objektmenge O (Universum), Distanzfunktion d: $O \times O \rightarrow \Re_0^+$, Datenbank DB $\subseteq O$

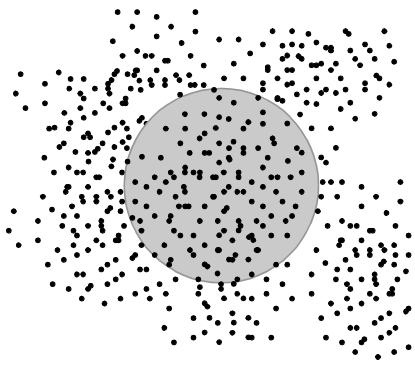
Bereichsanfragen

- Anfrageparameter: Anfrageobjekt q, maximaler Ähnlichkeitsabstand ε
- Ergebnismenge: $\text{sim}_\varepsilon(q) = \{ o \in \text{DB} \mid d(o, q) \leq \varepsilon\}$
- Anzahl der Ergebnisse: im vorhinein unbekannt, zwischen 0 und |DB|
- Ergebnisbereich: spezifizierter Bereich ε

Problem der Bereichsanfragen: Wie groß soll ε gewählt werden?



ε zu klein: keine Ergebnisse

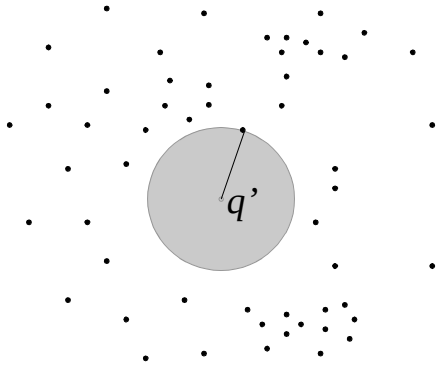


ε zu groß: zu viele Ergebnisse

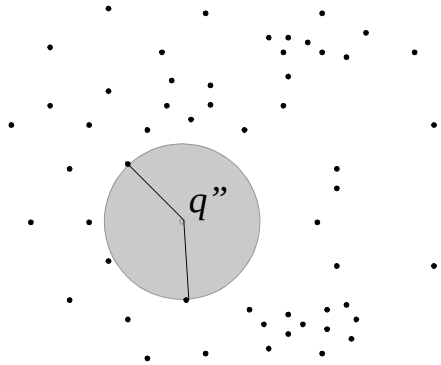
Nächste-Nachbarn-Anfragen

- Anfrageparameter: nur Anfrageobjekt q
- Ergebnismenge: $NN(q) = \{ o \mid \forall o' \in \text{DB}: d(o, q) \leq d(o', q)\}$
- Anzahl der Ergebnisse: 1 (mindestens) — auch Definition für "genau 1" möglich
- Ergebnisbereich: im vorhinein unbekannt, $\varepsilon_1 = \min \{d(o, q) \mid o \in \text{DB}\}$

Illustration:



eindeutiger nächster Nachbar

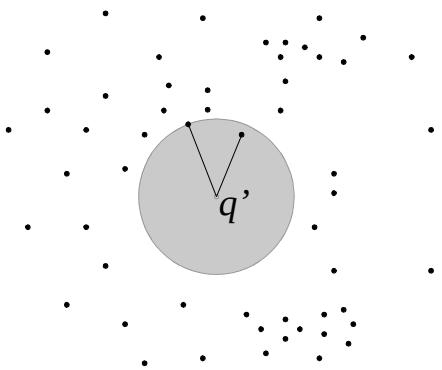


mehrere nächste Nachbarn

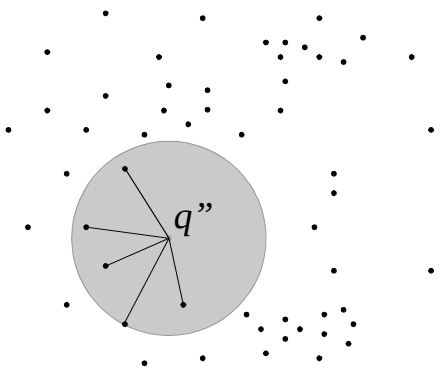
k-nächste-Nachbarn-Anfragen

- Anfrageparameter: Anfrageobjekt q, Anzahl gewünschter Ergebnisse k
- Ergebnismenge: kleinste Menge $NN_q(k) \subseteq \text{DB}$ mit $|NN_q(k)| \geq k$ für die gilt: $\forall o \in NN_q(k): \forall o' \in \text{DB} - NN_q(k): \; d(o, q) < d(o', q)$
- Anzahl der Ergebnisse: k (mindestens)
- Ergebnisbereich: im vorhinein unbekannt, $\varepsilon_k = \max \{d(o, q) \mid o \in NN_q(k)\}$

Beispiele:



$k = 2$



$k = 5$

Inkrementelles Ranking (*Give-me-more* Query)

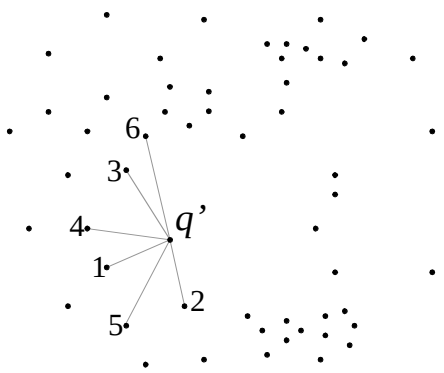
- Motivation
  - Oft kennt man weder brauchbare ε noch vernünftige k zu Beginn einer Recherche
  - Beispiel: Internet-Suchmaschinen
  - Gewünscht ist eine sortierte Ausgabe nach Abstand zum Anfrageobjekt

- Ablauf
  - Spezifikation eines Anfrageobjektes q beim Start.
  - Wiederholte Aufrufe der Funktion *getnext*(k_i), die jeweils die nächsten k_i Ergebnisse liefern, bis die gewünschte Ergebnismenge erreicht ist.
  - Es wird also schrittweise für eine aufsteigende Folge $K_1, K_2, \ldots$ mit $K_n = \sum_{i=1}^{n} k_i$ die Menge $NN_q(K_n)$ bestimmt (hier: jeweils genau K_n Elemente, auch bei gleichem Abstand nicht mehr).
  - Der Inhalt der Datenbank wird also (partiell) aufgezählt, und zwar aufsteigend nach dem Abstand zum Anfrageobjekt, d.h. für zwei Objekte o_i und o_j in dieser Aufzählung gilt:

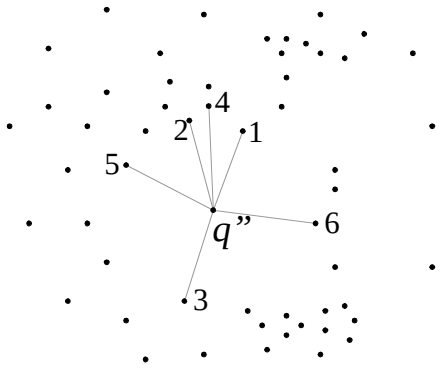
$$\forall\, i, j \in \{1, \ldots, N\}: \quad i < j \Rightarrow \mathrm{d}(o_i, q) \leq \mathrm{d}(o_j, q)$$

- Charakterisierung
  - Anfrageparameter: Anfrageobjekt q, Aufrufe von *getnext*(k_i)
  - Ergebnismenge: $NN_q(k)$ mit $k = \sum_{i=1}^{n} k_i$ für n Aufrufe von *getnext*(k_i)
  - Anzahl der Ergebnisse: $k = \sum_{i=1}^{n} k_i$ für n Aufrufe von *getnext*(k_i)
  - Ergebnisbereich: im vorhinein unbekannt, $\varepsilon_k = \max\{\mathrm{d}(o, q) \mid o \in NN_q(k)\}$

- Beispiele



6x *getnext*(1) um q'



6x *getnext*(1) um q"

2.1.2 Bewertung von Methoden zur Ähnlichkeitssuche

Übersicht

	erwünscht	unerwünscht
gefunden	richtig positive	falsch positive
nicht gefunden	falsch negative	richtig negative

- Begriffspaar Recall / Precision (aus Information Retrieval):
  - *Recall:* Wieviele der erwünschten Objekte wurden gefunden?
    rp / (rp + fn) = gefundene erwünschte Objekte / alle erwünschten Objekte
  - *Precision:* Wieviele der gefundenen Objekte sind erwünscht?
    rp / (rp + fp) = gefundene erwünschte Objekte / alle gefundenen Objekte

- Begriffspaar Sensitivität / Spezifität (aus Statistik):
  - *Sensitivität:* Wahrscheinlichkeit, daß Test für eine wahre Statistik positiv verläuft.
    rp / (rp + fn) = richtig positive / alle erwünschten Objekte (= recall)
  - *Spezifität:* Wahrscheinlichkeit, daß Test für eine falsche Statistik negativ verläuft.
    rn / (rn + fp) = richtig negative / alle unerwünschten Objekte

2.1.3 Klassen von Distanzfunktionen

- positiv-semidefinite Distanzfunktionen:
  $d(p, q) \geq 0$ (d.h. $d(p, q) = 0$ für $p \neq q$ möglich).
- positiv-definite Distanzfunktionen:
  $d(p, q) > 0$ für $p \neq q$, d.h. $d(p, q) = 0$ genau für $p = q$.
- Metriken:
  (*i*) Symmetrisch: $d(p, q) = d(q, p)$
  (*ii*) Definit: $d(p, q) = 0$ gdw. $p = q$
  (*iii*) Dreiecksungleichung: $d(p, q) \leq d(p, o) + d(o, q)$

Beispiele für Distanzfunktionen in n-dimensionalen Vektorräumen:

- Allgemeine L_p-Distanz: $d(o, q) = \left(\sum_{i=1}^{n} |o_i - q_i|^p \right)^{1/p}$
- $p = 2$, euklidischer Abstand: $d(o, q) = \sqrt{(o - q)^2}$
- $p = \infty$, Maximumsabstand: $d(o, q) = \max \{ |o_i - q_i|, i = 1, \ldots, n \}$
- $p = 1$, Summenabstand, "Manhattandistanz": $d(o, q) = \sum_{i=1}^{n} |o_i - q_i|$
- Gewichtete L_p-Distanzen: Benutzer kann Gewichte ändern
- Quadratische Formen: $d_A(o, q) = \sqrt{(o - q) \cdot A \cdot (o - q)^T}$ mit Ähnlichkeitsmatrix A