Skript zur Vorlesung

Datenbanksysteme II

Sommersemester 2005

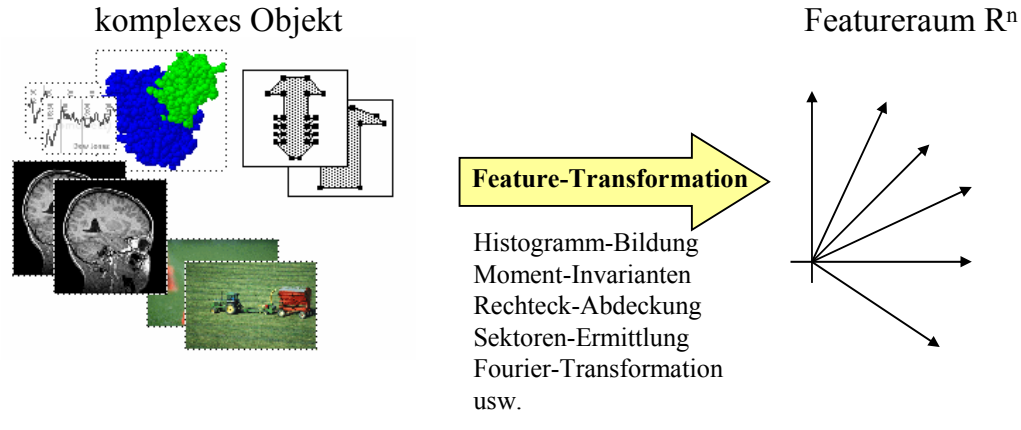
Kapitel 5: Einführung in Multimedia-Datenbanken

Vorlesung: Christian Böhm
Übungen: Elke Achtert, Peter Kunath

http://www.dbs.informatik.uni-muenchen.de/Lehre/DBSII

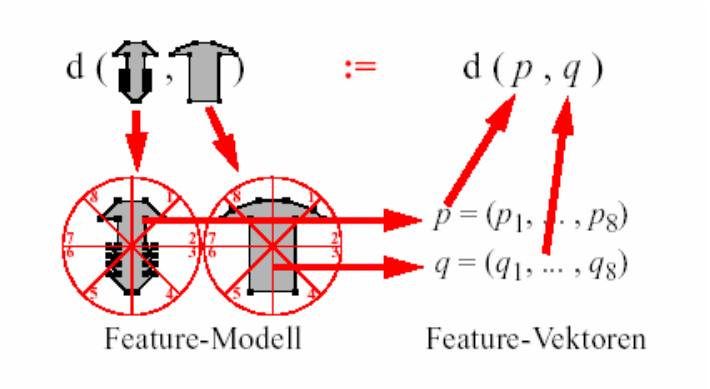
Featurebasierte Ähnlichkeit (1)

- **Feature-Transformation**
  - Kodierung komplexer Objekte durch hochdimensionale Vektoren
  - Extraktion numerischer Features aus den Objekten, die die Objekte charakterisieren → Feature-Vektoren



Featurebasierte Ähnlichkeit (2)

- **Distanzbasiertes Ähnlichkeitsmaß**
  - Ähnlichkeit der Objekte ist *unmittelbar* durch den Abstand der zugeordneten Featurevektoren *definiert*
  - Abstand im Featureraum steht für Unähnlichkeit der Objekte



Klassen von Distanzfunktionen

- **Positiv-semidefinite Distanzfunktion**
  $d(x, y) \geq 0$ (d.h. $d(x, y) = 0$ für $x \neq y$ möglich)

- **Positiv-definite Distanzfunktion**
  $d(x, y) > 0$ für $x \neq y$, d.h. $d(x, y) = 0$ genau für $x = y$.

- **Metrik**
  (*i*) Symmetrisch: $d(x, y) = d(y, x)$
  (*ii*) Definit: $d(x, y) = 0 \Leftrightarrow x = y$
  (*iii*) Dreiecksungleichung: $d(x, z) \leq d(x, y) + d(y, z)$

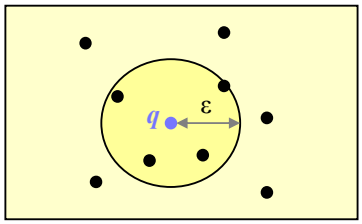
Beispiele für Distanzfunktionen

- **Allgemeine L_p-Distanz**: $d(x,y) = \sqrt[p]{\sum_{i=1}^{d} |x_i - y_i|^p}$
- **Euklidischer Abstand** (p=2): $d(x,y) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$
- **Maximumsabstand** (p=∞): $d(x,y) = \max\{|x_i - y_i|, i = 1 \ldots d\}$
- **Manhattandistanz** (p=1): $d(x,y) = \sum_{i=1}^{d} |x_i - y_i|$
- **Gewichtete L_p-Distanzen**: Benutzer kann Gewichte ändern
- **Quadratische Formen**: (mit Ähnlichkeitsmatrix A): $d(x,y) = \sqrt{(x-y) \cdot A \cdot (x-y)^T}$

Typen von Ähnlichkeitsanfragen (1)

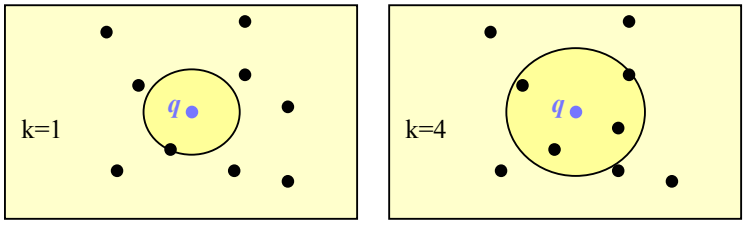
Basis: Objektmenge O, Distanzfunktion $dist: O \times O \rightarrow \Re_0^+$, Datenbank $DB \subseteq O$

- **Bereichsanfrage**
  - Anfrageparameter: Anfrageobjekt q, max. Ähnlichkeitsabstand ε
  - Ergebnismenge: $sim_\varepsilon(q) = \{o \in DB \mid d(o,q) \leq \varepsilon\}$
  - Anzahl Ergebnisse: im vorhinein unbekannt, zwischen 0 und $|DB|$
  - Ergebnisbereich: spezifizierter Bereich ε



Typen von Ähnlichkeitsanfragen (2)

- **Nächste-Nachbarn-Anfrag**
  - Anfrageparameter: Anfrageobjekt q
  - Ergebnismenge: $NN(q) = \{o \mid \forall o' \in DB : d(o,q) \leq d(o',q)\}$
  - Anzahl Ergebnisse: mind. 1
  - Ergebnisbereich: im vorhinein unbekannt, $\varepsilon_1 = \min\{d(o,q) \mid o \in DB\}$
- ***k*-Nächste-Nachbarn-Anfrage**
  - Anfrageparameter: Anfrageobjekt q,
    Anzahl gewünschter Ergebnisse k
  - Ergebnismenge: kleinste Menge $NN_q(k) \in DB$ mit $|NN_q(k)| \geq k$ für die gilt :
    $\forall o \in NN_q(k) : \forall o' \in DB - NN_q(k) : d(o,q) < d(o',q)$
  - Anzahl Ergebnisse: mind. k
  - Ergebnisbereich: im vorhinein unbekannt, $\varepsilon_k = \max\{d(o,q) \mid o \in NN_q(k)\}$



Typen von Ähnlichkeitsanfragen (3)

- **Inkrementelles Ranking (*Give-me-more* Query)**
  - 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.
    - Der Inhalt der DB wird also (partiell) aufgezählt, und zwar aufsteigend nach dem Abstand zum Anfrageobjekt, d.h. für zwei Objekte o_i und o_j gilt:
      $\forall i, j \in \{1,\ldots N\} : i < j \Rightarrow d(o_i,q) \leq d(o_j,q)\}$
  - 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 Ergebnisse: $k = \sum_{i=1}^{n} k_i$ für n Aufrufe von *getnext*(k_i)
  - Ergebnisbereich: im vorhinein unbekannt, $\varepsilon_k = \max\{d(o,q) \mid o \in NN_q(k)\}$

Bewertung von Methoden zur Ähnlichkeitssuche

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

- **Recall**: Wie viele der erwünschten Objekte wurden gefunden?

$$\frac{rp}{rp+fn} = \frac{\text{gefundene erwünschte Objekte}}{\text{alle erwünschten Objekte}}$$

- **Precision**: Wie viele der gefundenen Objekte sind erwünscht?

$$\frac{rp}{rp+fp} = \frac{\text{gefundene erwünschte Objekte}}{\text{alle gefundenen Objekte}}$$

- **Sensitivität**: WS, dass Test für gewünschtes Obj. positiv verläuft (=Recall)

$$\frac{rp}{rp+fn} = \frac{\text{richtig positiv}}{\text{alle erwünschten Objekte}}$$

- **Spezifität**: WS, dass Test für unerwünschtes Obj. negativ verläuft

$$\frac{rn}{rn+fp} = \frac{\text{richtig negativ}}{\text{alle unerwünschten Objekte}}$$

Mehrstufige Anfragebearbeitung

Qualitätskriterien für Filter

- **Eigene Effizienz**
  - Indextauglichkeit
  - schnelle Auswertung
- **Vollständigkeit**
  - Kandidaten müssen alle Ergebnisse enthalten
  - Beweis durch Lower-Bounding-Lemma: $d_{filter}(p,q) \leq d_{exact}(p,q)$
  - Analog gilt bei mehreren Filtern: $d_{filter1}(p,q) \leq d_{filter2}(p,q) \leq \ldots$
- **Gute Selektivität**
  - Möglichste wenig Kandidaten für Verfeinerung
  - datenabhängig → empirischer Nachweis