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

Skript zur Vorlesung

Datenbanksysteme II

Sommersemester 2006

Kapitel 8: Ähnlichkeitsmodelle für Polygone und 3D Daten

Vorlesung: Christian Böhm
Übungen: Elke Achtert, Peter Kunath, Alexey Pryakhin

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

Datenbanksysteme II
Kapitel 8: Ähnlichkeitsmodelle für Polygone und 3D Daten

Inhalt

Formhistogramme für 3D-Objekte

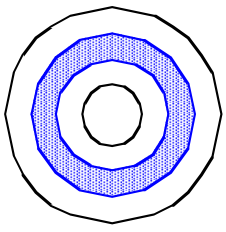
([AKKS 99] Ankerst M., Kastenmüller G., Kriegel H.-P., Seidl T.: 3D Shape Histograms for Similarity Search and Classification in Spatial Databases. Proc. Int. Symposium on Large Spatial Databases (SSD) 1999 (LNCS 1651), 207-226.)

- **Ziel**
  - Translations- und rotationsinvariante Suche nach ähnlichen Formen im 3D.
  - Objekte sind als Mengen von Oberflächenpunkten gegeben.
  - Beispielanwendungen: Moleküle, CAD-Bauteile.

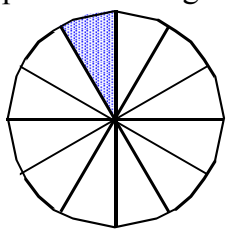
- **Grundidee: Formhistogramme**
  - Partitioniere den 3D-Raum in Zellen (Histogramm-Bins).
  - Bestimme den Anteil an Punkten des Objektes pro Zelle (normiertes Histogramm).
  - Durch die Normierung werden die Histogramme unabhängig von der Punktedichte.

Formhistogramme, Beispiele

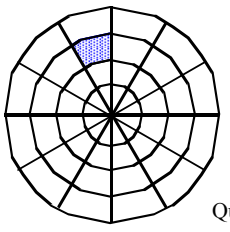
- Verschiedene Raumpartitionierungen



Schalenmodell



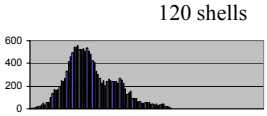
Sektorenmodell



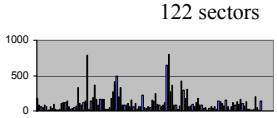
kombiniertes Modell

Quelle: [AKKS 99]

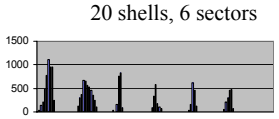
- Beispielobjekt Seryl-tRNA Synthetase (PDB-Code: 1SER-B)



Schalenmodell



Sektorenmodell



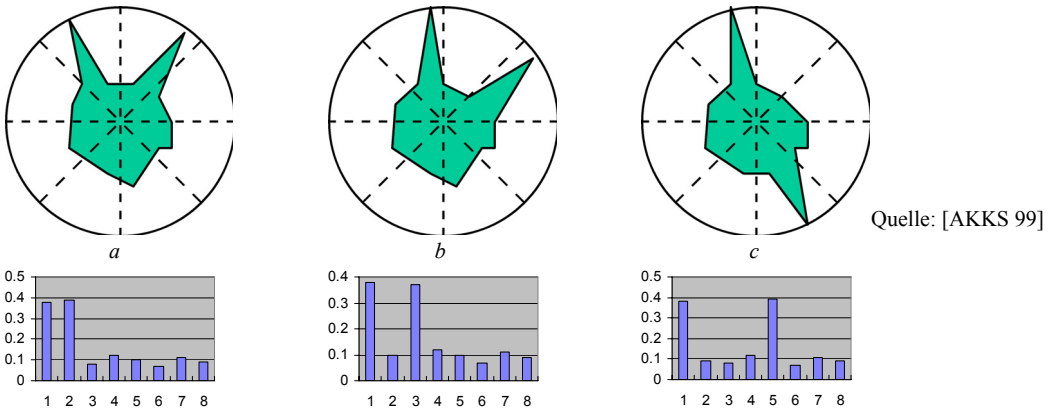
kombiniertes Modell

Formhistogramme, Definition

- **Formale Definition der Histogramme**
  - *Schalenmodell*: Definiere die Bins über den Abstand zum Mittelpunkt, d.h. Anzahl der Punkte auf der jeweiligen Schale.
  - *Sektorenmodell*: Anzahl der Punkte im jeweiligen Sektor.
  - *Kombiniertes Modell*: Synthese aus Schalen- und Sektorenmodell.
- **Invarianzen**
  - Translationsinvarianz durch Lagenormierung: Verschiebung des Schwerpunkts eines Objektes in den Ursprung.
  - Rotationsinvarianz durch Hauptachsentransformation:
    - Drehung der Objekte, so dass die Hauptachsen auf den Koordinatenachsen liegen.
    - unnötig beim Schalenmodell, dieses ist inhärent rotationsinvariant.

Formhistogramme, Distanzfunktion (1)

- Probleme mit dem euklidischen Abstand (Beispiel im 2D)



Quelle: [AKKS 99]

  - Die Form *c* gilt als genauso ähnlich zu *a* wie zu *b*.
  - Die Ähnlichkeit räumlich benachbarter Histogramm-Bins wird nicht berücksichtigt.

Formhistogramme, Distanzfunktion (2)

- **Quadratische Formen als Distanzfunktionen**

$$d_A(p,q) = \sqrt{(p-q)\cdot A\cdot (p-q)^T} = \sqrt{\sum_i \sum_j a_{ij}\cdot (p_i - q_i)\cdot (p_j - q_j)}$$

  - Für die Formhistogramme enthält die Ähnlichkeitsmatrix $A = [a_{ij}]$ die Ähnlichkeit von Einträgen in den Zellen i und j der Raumpartitionierung
  - Diese Ähnlichkeit läßt sich aus dem Abstand d_{ij} der Zellen i und j berechnen, z.B.:
    $a_{ij} = \exp(-\sigma\,(d_{ij} / d_{max})^2)$
  - Als Abstand d_{ij} eignet sich beispielsweise der euklidische Abstand

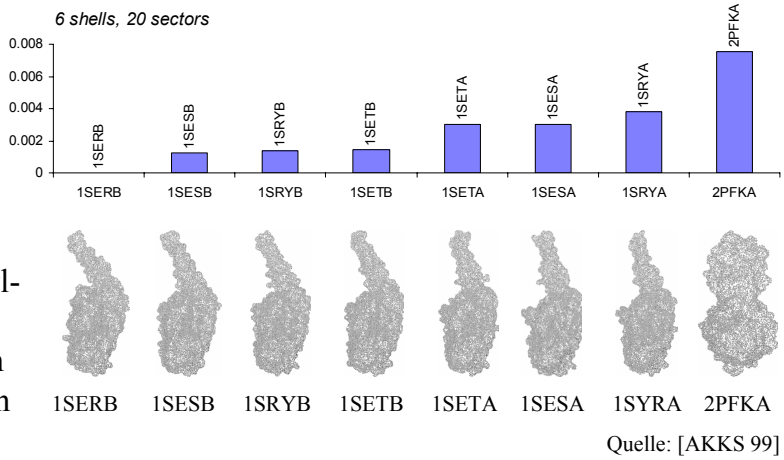
Formhistogramme, Ergebnisse (1)

- **Einfache Suche nach ähnlichen Molekülen**

Anfrage: 1SER-B

Die erwarteten Ergebnisse (Seryl-Proteasen) treten auf den ersten Positionen auf.

Das erste nicht-Seryl-Protein (2PFKA) unterscheidet sich in seiner Form sowie in seinem Distanzwert zur Anfrage deutlich von den Seryl-Proteinen.



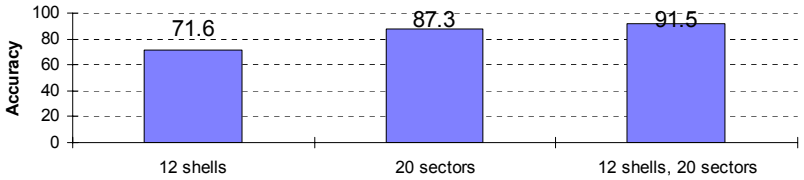
Quelle: [AKKS 99]

Formhistogramme, Ergebnisse (2)

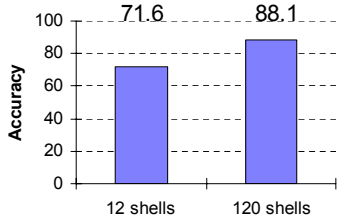
- **Klassifikation auf der gesamten Datenbank**
  - Das Ähnlichkeitsmodell wird in einem nächsten-Nachbar-Klassifikator verwendet.
  - D.h. ein Anfrageobjekt bekommt das Klassenlabel des ähnlichsten Objektes aus der Datenbank zugeordnet.
  - Klassifikationsgenauigkeit: Wie oft wird die Klassenentscheidung richtig getroffen?
  - "Leave-One-Out"-Experiment: jedes Objekt wird gegen die restliche DB angefragt.

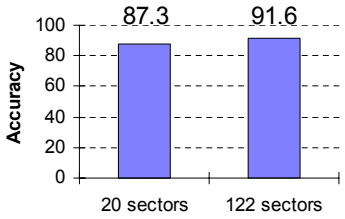
Formhistogramme, Ergebnisse (3)

- **Ergebnisse**
  - Vergleich verschiedener Histogrammodelle (12 Schalen, 20 Sektoren, 12×20 Zellen).



  - Vergleich verschieden-granularer Raumpartitionierungen (12 bzw. 120 Schalen bzw. 20 bzw. 122 Sektoren).

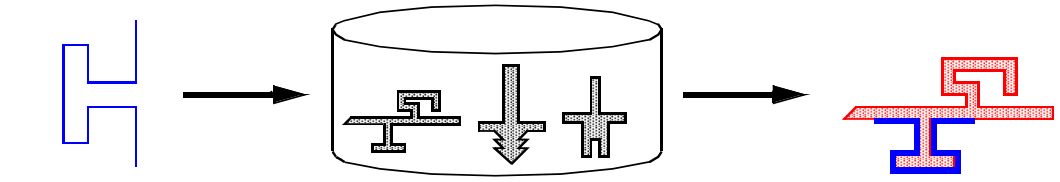




Partielle Ähnlichkeitssuche (1)

([BKK 97] Berchtold S., Keim D. A., Kriegel H.-P.: *Using Extended Feature Objects for Partial Similarity Retrieval.* VLDB Journal 6(4), 1997, 333-348.)

- **Ziel**
  - Translations-, rotations- und skalierungsinvariante Ähnlichkeit von Polygonen.
  - Unterstützung von partieller Ähnlichkeitssuche.

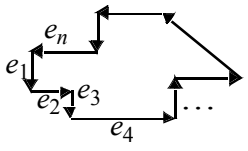


Anfragekantenzug | DB von Polygonen | partiell ähnliches Polygon

Partielle Ähnlichkeitssuche (2)

- **Polygondarstellung**
  - Ein Polygon ist ein geschlossener Kantenzug $\langle e_1, \ldots, e_n \rangle$.
  - Polygon wird als parametrische Kurve $p(t)$ dargestellt, wobei t von 0 bis 2π läuft.
  - Da der Kantenzug geschlossen ist, gilt:

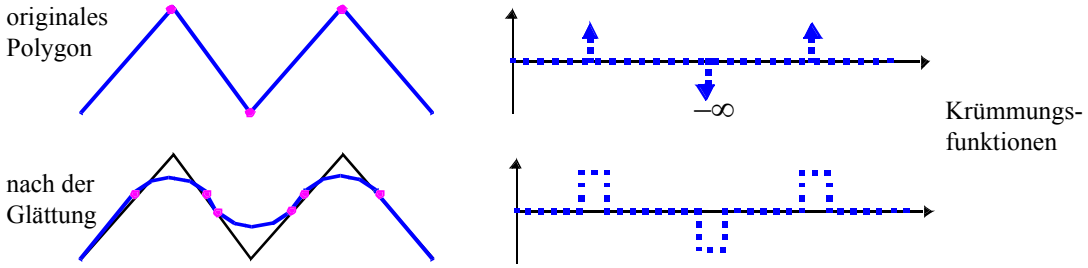
$$p(2\pi) = \int_{0\ldots2\pi} p(t)dt = \sum_{i=1}^{n} e_i = p(0) = 0$$



  - Die Länge des Kantenzuges ist auf 2π normiert: $\sum_{i=1}^{n} |e_i| = 2\pi$
  - Über zwei Parameter a und b lassen sich Ausschnitte aus Polygonen beschreiben.
- **Invariante Darstellungen**
  - *Skalierungsinvarianz:* Durch Normierung auf Länge 2π erreicht.
  - *Translationsinvarianz*: Die Polygone sind nicht absolut positioniert.
  - *Rotationsinvarianz:* Betrachte statt der konkreten Richtungen die Krümmungen!
  - Invarianz gegenüber gewähltem Startpunkt: In partieller Ähnlichkeit enthalten

Partielle Ähnlichkeitssuche (3)

- **Repräsentation der Krümmungen**
  - Problem: Kantenzug ist nicht stetig differenzierbar (scharfe Knicke an den Ecken).
  - Lösung: Glättung durch Approximation der Ecken mit Kreisausschnitten.



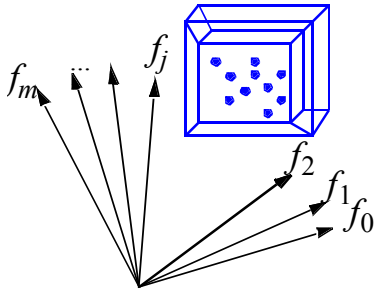
  - Die Amplituden der Krümmungsfunktion hängen vom gewählten Radius r ab.
  - Die Breite eines Ausschlages in der Krümmungsfunktion hängt vom Winkel α_i ab.
  - Für die Krümmungsfunktion eines Polygons werden die Fourier-Koeffizienten gespeichert (analytische Berechnung)

Behandlung der partiellen Ähnlichkeit (1)

- **Repräsentation von Polygonausschnitten.**
  - Ausschnitte (a, b) aus Kantenzügen $p(t)$, d.h. $a \leq t \leq b$, werden selbst wieder auf den Bereich 0, …, 2π skaliert. Der Datenraum ist zyklisch für den Fall $b < a$.
  - Für die Polygonausschnitte werden ebenfalls die Krümmungsfunktionen berechnet und deren Fourier-Koeffizienten ermittelt.
  - Für den Featureraum werden einige der Koeffizienten als Dimensionen ausgewählt.
  - Die Ähnlichkeit von Polygonausschnitten wird über eine geeignete Distanzfunktion im hochdimensionalen Featureraum definiert (z.B. p-Norm).

Behandlung der partiellen Ähnlichkeit (2)

- **Unendlichkeitsproblem**
  - Es gibt unendlich viele Ausschnitte (a, b) und damit Punkte im hochdimensionalen Featureraum, die weder alle berechnet noch alle gespeichert werden können.
  - Lösung: Speichere nicht die einzelnen Punkte, sondern jeweils das minimal umgebende Hyperrechteck mehrerer Featurepunkte.
  - Dazu verschiedene Strategien, wie viele Featurepunkte zusammengefasst werden.



Behandlung der partiellen Ähnlichkeit (3)

- **Zusammenfassen von Featurepunkten**
  - Erster Schritt: Zusammenfassen der Featurepunkte für Abschnitte, die auf derselben Kante des Polygons beginnen und auf einer bestimmten anderen Kante enden (d.h. $n \cdot n$ viele Hyperrechtecke).
  - Beobachtung: Manche dieser Boxen sind sehr klein, andere sehr groß.
  - Kleine Boxen können weiter zusammengefasst werden:
    - Die Hyperrechtecke für benachbarte Kanten im Polygon werden zusammengefasst.
  - Große Boxen können weiter zerlegt werden:
    - Zerlegung im Featureraum:
      - Auswahl bestimmter Achsen oder
      - Auswahl aller Achsen: dann gibt es 2^m viele Zerlegungsprodukte.
    - Alternative: Zerlege Boxen im zweidimensionalen Parameterraum, d.h. fasse andere Polygonausschnitte zusammen.