
Kapitel 1

Einführung

Skript zur Vorlesung: Spatial, Temporal, and Multimedia Databases
Sommersemester 2013, LMU München

© 2006 Prof. Dr. Hans-Peter Kriegel, Dr. Peer Kröger, Dr. Peter Kunath, Dr. Matthias Renz, Arthur Zimek

Vorlesung (Plenum):



PD Dr. Matthias Renz
Oettingenstr. 67, Zimmer F 111
Tel. 089/2180-9331
Sprechstunde: Dienstag, 13⁰⁰-14⁰⁰

Übungsbetrieb:



Markus Mauder
Oettingenstr. 67, Zimmer F 104
Tel. 089/2180-9321

Tutor:

Diana Uskat

Termine

- Vorlesung: Freitag 9:30 - 12 Uhr Raum L 155 (Oettingenstr. 67)
(Start: 19.04.2013)
- Übung: Montag, 14-16 Uhr Raum A 105 (Amalienstr. 17)
 Montag, 16-18 Uhr Raum A 105 (Amalienstr. 17)
(Start des Übungsbetriebs: 29.04.2012)

Anmeldung für den Übungsbetrieb auf der Homepage

[http://www.dbs.ifi.lmu.de/cms/Spatial, Temporal and Multimedia Databases](http://www.dbs.ifi.lmu.de/cms/Spatial,_Temporal_and_Multimedia_Databases)

Schein-/Punkteerwerb

- Zulassung: Anmeldung für den Übungsbetrieb
(über Uniworx auf der Homepage)
- Prüfung: Klausur

Inhalt der Vorlesung

- Ähnlichkeits-, Nachbarschafts- und Schnittanfragen in Datenbanken mit allg. komplex-strukturierten Daten:
 - aus der Wissenschaft (Medizin, Biologie, Archeologie, ...)
 - Multimedia Objekte (Bilder, Videos, ...)
 - Räumliche Objekte (CAD Daten)
 - Zeitliche (zeitabhängigen) Objekte bzw. Sequenzdaten (Temperaturkurven, Aktienkurse, ...)
- Ähnlichkeitsmodelle
 - Modellierung der Ähnlichkeit von räumlichen, zeitlichen und Multimedia Objekten
- Algorithmen zur Ähnlichkeits- bzw. Nachbarschaftssuche
 - Effiziente Bearbeitung von Nachbarschaftsanfragen

1.1 Motivierende Beispiele

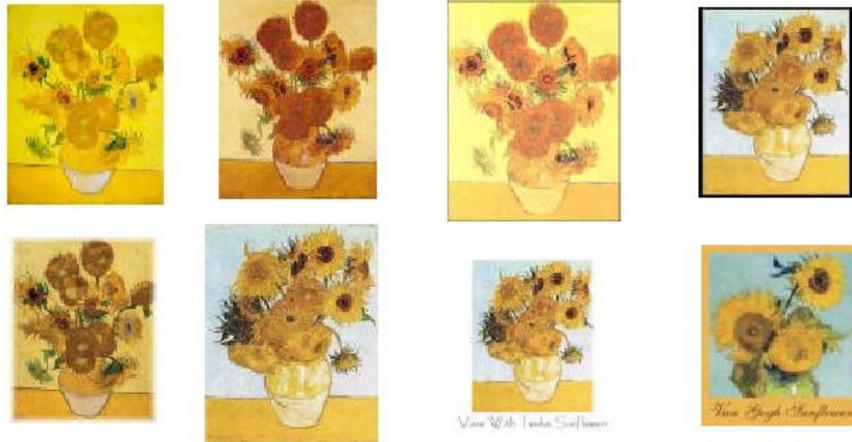
Beispiel 1: Bildersuche

- Gegeben: Archiv mit 2 Mio. Bildern (2D Objekte)
- Frage: Ist im Archiv ein bestimmtes Kunstwerk abgebildet?



– Herausforderung

- „abgebildet“ heißt nicht „identische Binärrepräsentation“ wie das Anfragebild



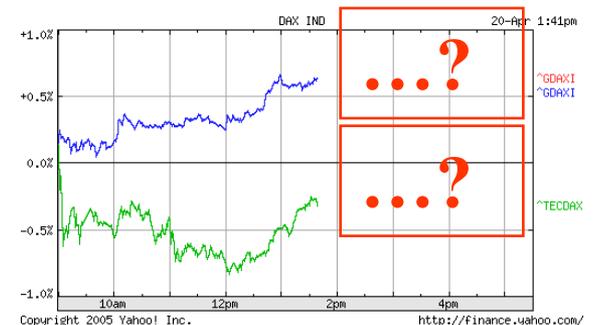
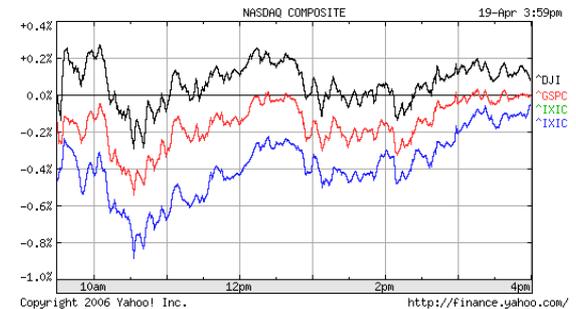
- Abweichungen

- Unterschiedliche Größe (Skalierung, Auflösung)
- Unterschiedliche Ausrichtung z.B. durch unterschiedliche Perspektive (Spiegelung, ...)
- Unterschiedliche Farbgebung (Tönung der Farben)
- Abweichende Ausschnittsbildung
- Hinzugefügter Rand oder Beschriftung
- ...

Beispiel 2: Trendanalyse für Aktienkurse

- Gegeben: Datenbank von Aktienkursen, Anfrage-Kurs
- Frage: Finde Aktien in der DB, die einen ähnlichen Kurs wie die Anfrage haben (um zukünftiges Verhalten vorherzusagen)

- Herausforderungen:
 - Zeitverschiebungen
 - Ausreißer
 - Unterschiedliche Skalierung
 - ...



Beispiel 3: Sequenzsuche in Videodateien

- Gegeben: Datenbank von Videofilmen, Abfragesequenz
- Gesucht: alle Videos in der DB, die eine Sequenz ähnlich der Abfragesequenz enthalten
- Herausforderungen:
 - Erkennen von Bildinhalten statt reiner Bildähnlichkeit
 - Unvollständige Bildsequenzen
 - Unterschiedlich lange Sequenzen
 - Unterschiedliche „Auflösung“ (FPS, Bildauflösung)
 - ...

Beispiel 4: Location-Based Services (LBS)

Definition: *Standortbezogene Dienste (Location-Based Services) = Dienste, die unter Zuhilfenahme von positions-, zeit- und personenabhängigen Daten dem Endbenutzer selektive Informationen bereitstellen* [Wikipedia]

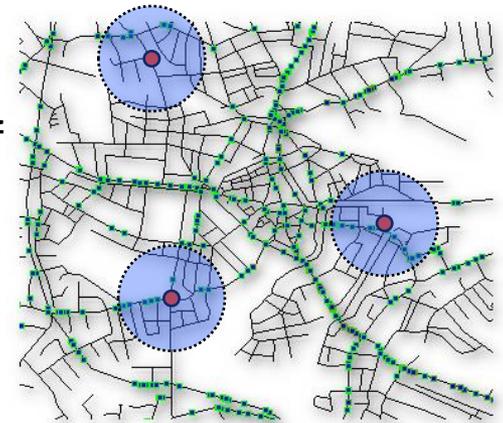
– Gegeben:

- Datenbank von sich bewegenden Objekten (z.B. Autos in einem Strassennetz)
- Menge von ausgezeichneten Positionen (z.B. Orte von Interesse wie z.B. Restaurants, etc.)

– Gesucht: alle Autos die sich in der unmittelbaren Nähe von Tankstellen der Firma „Tankgut“ befinden

– Herausforderungen:

- Indexierung von und Nachbarschaftsanfragen auf Objekten in nicht-Euklidischen Räumen
- Distanzberechnungen in Straßennetzwerken sind teuer
- ...



- Allgemeine Problemstellungen bei der Ähnlichkeitssuche
 - Informelle Ebene
 - Ähnlichkeit situationsabhängig, z.B. Bildsuche
 - » Suche nach „Abendrot“ => Farben wichtig
 - » Suche nach „Personen“ => Formen wichtig
 - Ähnlichkeit personenabhängig (z.B. rot/grün Blindheit)
 - Allgemein: Ähnlichkeit Gegenstand psychologischer Forschung
 - Formale Ebene
 - Mathematische Beschreibung von Objekten (Objektrepräsentation)
 - Mathematische Beschreibung der „Ähnlichkeit“ zum Vergleich von Objekten
 - Ähnlichkeitsmaß: quantitative Bewertung der Ähnlichkeit zweier Objekte durch eine Maßzahl (z.B. „100% ähnlich“); komplementär: Distanzmaß (z.B. „Abstand gleich 0“)
 - Pragmatische Ebene
 - (effizienter) Algorithmus zur Bestimmung der Ähnlichkeit zwischen zwei Objekten
 - (effizienter) Algorithmus zur Suche von ähnlichen Objekten in einer großen Datenbank

– Teilproblem der Suche

- Sequentielle Suche („sequential scan“)
 - Vergleich des Anfrageobjekts mit jedem einzelnen Datenbankobjekt
 - Skaliert *linear* zur Größe der Datenbank, d.h. 100-mal mehr Objekte => 100-mal längere Suchzeit
 - => für große Datenbanken dauert Suche „viel zu lange“
- Herausforderungen
 - Beschleunigung der Suche (geschickte Datenorganisation)
 - Beschleunigung der Einzelvergleiche (geeignete Repräsentationen)

– Lösungsansatz 1: Annahme einer Normalform

- Normalform: es gibt Stringdarstellung $s(v)$, $s(w)$ für jedes Objekt v, w , sodass $s(v) = s(w) \Leftrightarrow w$ stellt v dar
- Bewährte Suchtechniken skalieren gut für sehr große Datenbanken (Suchbaum, Hashverfahren)
- ABER: geeignete Normalform(en) schwierig (sehr unwahrscheinlich) zu finden

– Lösungsansatz 2: Feature-basierte Ähnlichkeit

- z.B. Bildsuche: Einfache Eigenschaft eines Bildes:
Durchschnittsfarbe

$$\text{avg: pic} \rightarrow (r, g, b)$$

dann gilt

$$v \text{ stellt } w \text{ dar} \Rightarrow \text{avg}(v) = \text{avg}(w)$$

Ähnlichkeitsanfrage: $|\text{avg}(v) - \text{avg}(w)| \leq \varepsilon$

- Sinnvoll, falls nicht zu viele Bilder ε -ähnlich (kleine Selektivität der Anfrage)
- Mehrstufiges Vorgehen: avg als Filter, genauer Vergleich als Verfeinerung

Mögliche Erweiterungen

- Farbhistogramme statt einfache Durchschnittsfarbe
- Beziehungen der einzelnen Dimensionen berücksichtigen
- Berücksichtigung der dargestellten Formen (geometrische Ebene)
- Berücksichtigung von dargestellten Objekten (semantische Ebene)
- Erweiterung auf Bildfolgen (Videos)
- ...

1.2 Inhalt der Vorlesung

1. Einführung
2. Prinzipien der Anfragebearbeitung in STMM-DBS
3. Räumliche Anfrage und Ähnlichkeitssuche für räumlich ausgedehnte Objekte
4. Ähnlichkeitssuche in zeitlichen Objekten & Sequenzen

1.3 Warum STMM-Datenbanken

- Spatial-, Temporal-, Multimedia-Objects
 - Irreversibler Trend in der IT
 - Neue Qualität von Informationen
 - Beschreibung von komplexen Strukturen, dynamischen Vorgängen, etc.
 - i.A. geringerer Informationsverlust wenn Verarbeitung im jeweiligen „Medium“ statt Umsetzung in ein anderes Medium

- Warum Datenbank-Technologie für STMM?
 - Sehr große Mengen an Daten vorhanden
 - Speicherplatzintensive Daten
 - In vielen STMM-Anwendungen ist Mehrbenutzerbetrieb erwünscht
 - Daten sollen (effizient und effektiv) recherchierbar sein

– Standard-DBS

- Konsistenzerhaltender Mehrbenutzerbetrieb
- Physische und logische Datenunabhängigkeit
- Effiziente Anfragebearbeitung durch geeignete Speicherungsstrukturen
- Unterstützung von Transaktionen
 - Concurrency: Isolation gleichzeitiger Updates verschiedener Benutzer
 - Recovery: konsistentes Wiederaufsetzen im Fehlerfall
 - Überwachung der Datenintegrität
- Datensicherheit, Datenschutz

– STMM-DBS

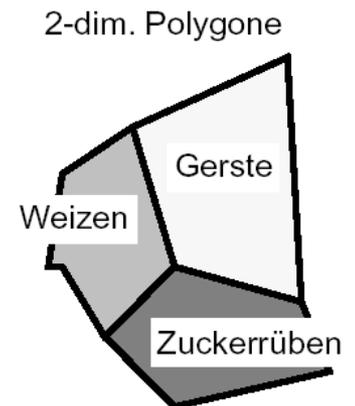
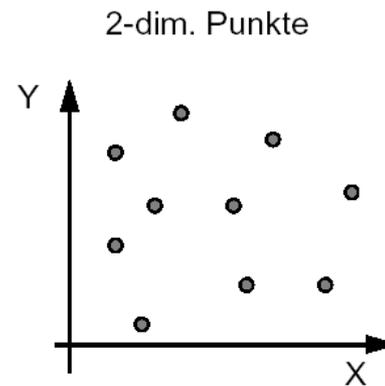
- Verschiedene Verwendungen des Begriffs, z.B. für:
 - CD Sammlungen, die Infos mittels Stichwortsuche zugreifbar machen
 - Systeme zur Organisation und Sichtung von Informationen mittels Browser (z.B. Wiki)
 - Video-on-demand-Systeme
 - CAD-Systeme, die DBS nutzen
 - Relationale DBS, die zusätzlich sog. BLOBs (Binary Large Objects) speichern
- In dieser Vorlesung ganz allgemein ein DBS
 - mit hoher Kapazität und Performanz
 - das Spatial-, Temporal-, und Multimedia-Datentypen sowie alphanumerische Datentypen unterstützt
 - das mit großen Datenvolumina umgehen kann

1.4 Anwendungen

1.4.1 Anwendungen mit räumlichen Daten

– Geoinformationssysteme (GIS)

- Geo-Objekte typischerweise 2D oder 3D mit räumlichen Bezugspunkt
- Oft zusätzliche räumliche und nicht-räumliche Attribute
 - Geometrische Attribute
(Koordinaten, Fläche, Umfang, ...)
 - Topologische Attribute
(Nachbarschaftsbeziehung, ...)
 - Thematische Attribute
(„Landuse“, ...)



- Datenmodellierung

- Vektormodell

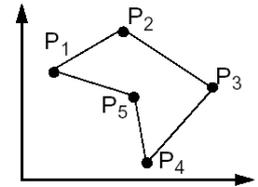
- » Geo-Objekte werden durch ihren Rand beschrieben
 - » Der Rand wird durch eine Menge von Punkten definiert

- Rastermodell

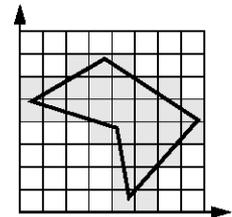
- » Geo-Objekte werden durch ihr Inneres beschrieben
 - » Das Innere wird als Menge von Pixeln eines Gitters definiert

Vektormodell:

$(P_1, P_2, P_3, P_4, P_5)$



Rastermodell:



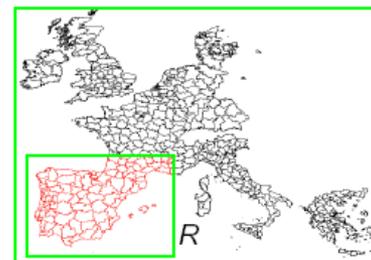
- Anfragen/Recherche

- Thematische Anfragen (“Suche alle Krankenhäuser”)
 - Geometrische Anfragen

Punkt-Anfrage

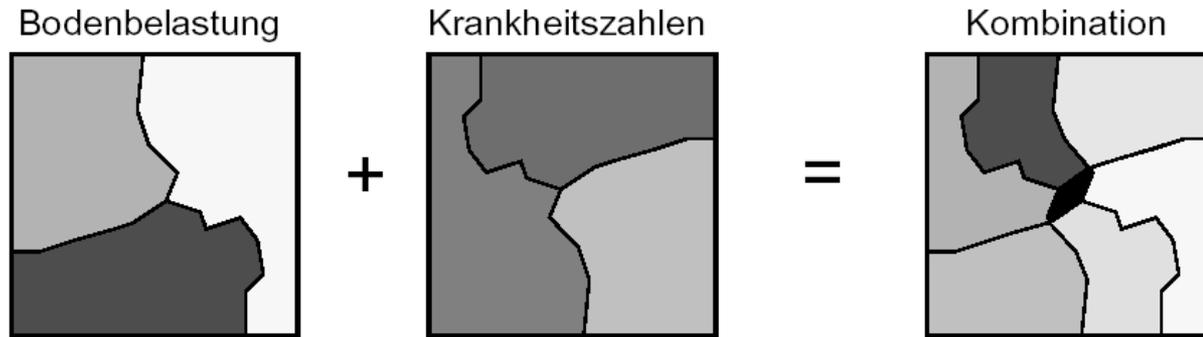


Fenster-Anfrage

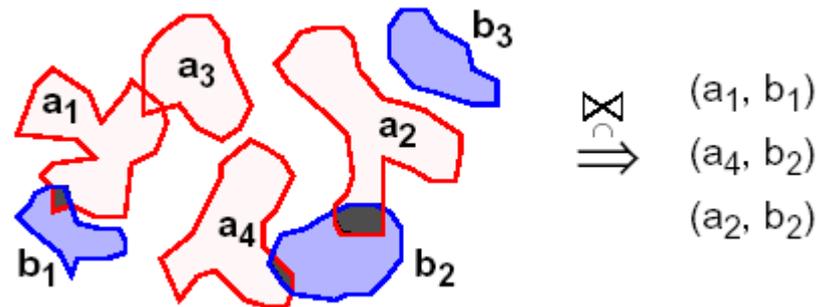


- Thematisch-geometrische Anfragen (“Suche alle Krankenhäuser in einem Umkreis von 50 km”)

- *Map Overlay*: Verschneidung von 2 oder mehr Karten (Kombination von Geometrie und Thematik)

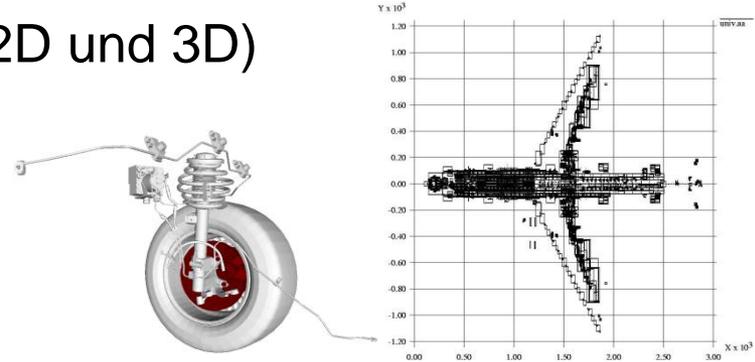


=> Unterstützung durch DBS: *Spatial Join*



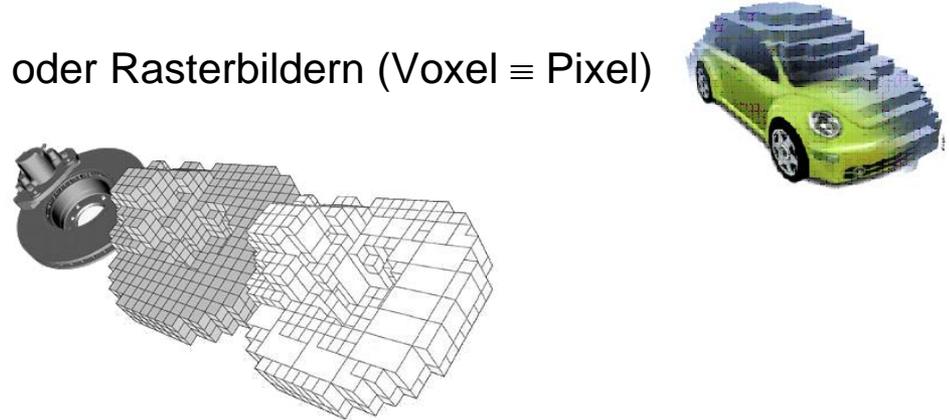
– CAD-Systeme, Digital Mock-up, Haptische Simulation

- Komplexe geometrische Objekte (2D und 3D)
 - Bauteil-Oberflächen und -Volumina
 - Hüllgeometrien



- Datenmodellierung

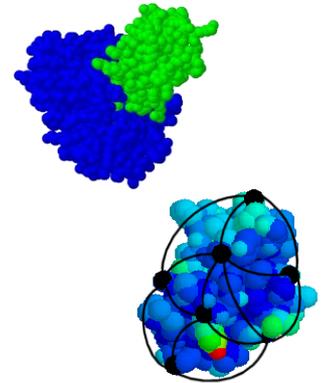
- 3D Erweiterung von Vektor- oder Rasterbildern (Voxel \equiv Pixel)
- Darstellung durch
 - » Triangulation
 - » Voxelisierung
 - » Intervalle
 - » ...



- Anfragen/Recherche

- ähnliche Formen
- Schnitt-/Kollisionsanfragen

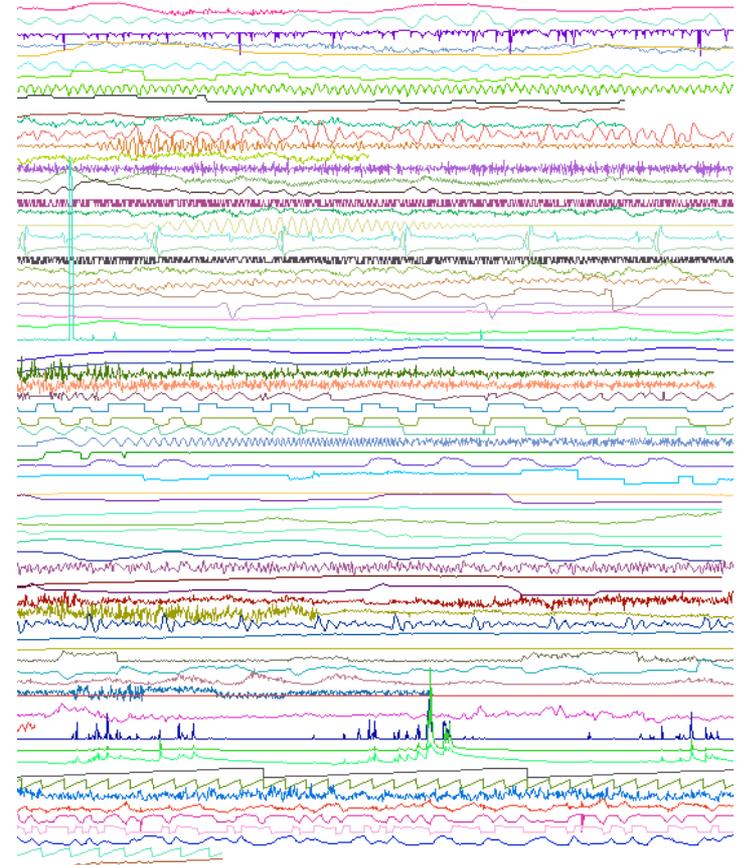
- Biomoleküle (Proteindocking, Medikamentenentwurf, ...)
 - 3D Objekte i.d.R. bestehend aus 3D Atomkoordinaten
 - Zusätzliche physikochemische Eigenschaften der Atome
 - Unterschiedlichste Datenmodellierungen
 - 3D Punktmengen (Atomkoordinaten)
 - Oberflächenmodellierungen (Triangulation, Voxelisierung)
 - Graphmodellierung
 - ...
 - Anfragen/Recherche
 - Geometrische Ähnlichkeit
 - Suche nach komplementärer Oberfläche (Dockingpartner)
 - Berücksichtigung physikochemischer Eigenschaften



1.4.2 Anwendungen mit Zeitreihen und Sequenzen

– Zeitreihenanalyse

- Meist periodisch aufgenommene Messwerte einer physikalischen Größe
 - Aktienkurse
 - Fieberkurven
 - Herzfrequenz
 - Sensoren
 - ...
- Datenmodellierung
 - Einfache Listen von Messwerten
- Anfragen/Recherche
 - Komplette Ähnlichkeit
 - » Ähnlichkeit in der Zeit (z.B. L_2 -Norm)
 - Partielle Ähnlichkeit
 - » Ähnlichkeit in der Gestalt (Muster)
 - » Autokorrelation



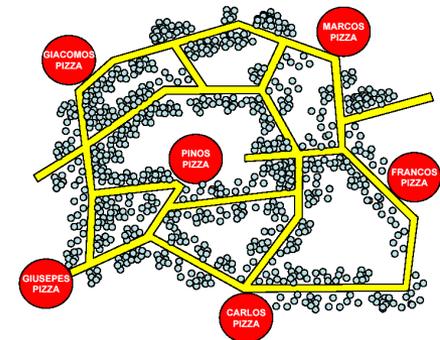
– Sequenzdaten

- Objekte sind Zeichenketten über ein (typischerweise fixes) Alphabet
 - Text optional erweitert um
 - » Strukturelle Informationen wie Titel, Autor, etc. z.B. durch Mark-up Sprachen (SGML, HTML, XML, ...)
 - » Layout-Informationen
 - Biologische Sequenzen
 - » Proteinsequenzen (3-Letter-Code, 1-Letter-Code)
 - » Nucleotidsequenzen (Alphabet meist aus den vier Zeichen {A,C,G,T})
- Anfragen/Recherche
 - Suche nach einzelnen Wörtern/Wortkombinationen (Volltextsuche)
 - Toleranz gegenüber Orthographie- oder Konvertierungsfehlern, Synonyme
 - Suche nach ähnlichen Dokumenten

1.4.3 Anwendungen mit beweglichen Objekten

– Analyse von Objekte in einem Verkehrsnetz

- Location-based Services, Verkehrsplanung und -monitoring
- Objekte, die sich entlang eines Verkehrsnetzes bewegen
 - Autos, Fußgänger, Züge, ...
- Modellierung
 - Verkehrsnetz \equiv (evtl. gerichteter) Graph
 - Knoten: Kreuzungen, ...
 - Kanten: Verbindungen (Straßen, ...)
 - Objekte auf Kanten oder Knoten platziert
- Anfragen/Recherche
 - Ähnlichkeit (bzgl. der Lage) zwischen Objekten über Netzwerkdistanz (Dijkstra-Algorithmus)
 - Suche nach räumlich nahen Objekten
 - Suche nach Objekten mit ähnlichen Bewegungspatterns



1.4.4 Typen von Multimediaobjekten

– Graphen

- Beschreiben viele Konzepte menschlichen Denkens (Schaltpläne, Organisationsstrukturen, ...)
- Anfragen/Recherche
 - Struktur des Graphen
 - Knotenattribute, Kantenattribute
 - Ähnlichkeitssuche nach
 - » Gesamtähnlichkeit (ganzer Graph)
 - » Partielle Ähnlichkeit (Teilgraphen)

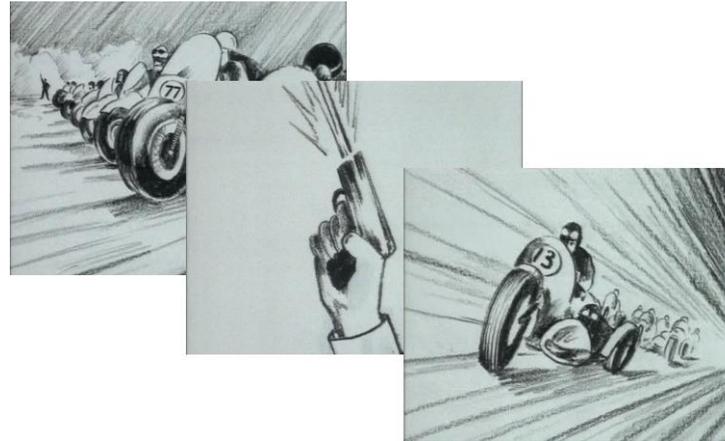
– Rasterbilder (still image)

- Modellierung
 - Zahlreiche Formate und Kompressionsverfahren (JPEG, GIF, PNG, TIFF, ...)
 - Matrix mit Bildpunkten (Pixel)
 - Pixeltiefe (Farbtiefe):
 - » 1 Bit pro Pixel => s/w-Bilder
 - » >1 Bit pro Pixel => Grauwert-/Farbbilder

- Farbtabellen
 - » Wenn wenige Bits für Codierung zur Verfügung stehen
 - » Fixe Bitanzahl z.B. für Grundfarben (RGB, CYM)
 - » Alle Farben lassen sich durch Mischung erzeugen
- Anfragen/Recherche
 - Suche nach ähnlichen Bildern (Farbe, Formen, ...)
- Audio
 - Sprache, Musik, ...
 - Modellierung:
 - Zeitreihen/Sequenzen, z.B. Pulse Code Modulation (PCM)
 - » Feste Zeitabstände (Abtastrate, sampling rate)
 - » Feste Bitanzahl zum Codieren (Abtasttiefe)
 - Musik oft in abstrakter Form (Notensequenz, MIDI-Daten)
 - Zahlreiche Kompressionsverfahren
 - Anfragen/Recherche
 - Mustererkennung mit oder ohne Berücksichtigung der Lautstärke/Sprechgeschwindigkeit etc.

– Video

- Sequenz von Rasterbildern oder Vektorgrafiken (Animation)



- Anfragen/Recherche
 - Ähnlichkeitssuche (Plagiaterkennung)
 - Subsequenzsuche
 - » Szenen
 - » Einstellungen
 - » ...

1.5 Recherche in STMM-DBS (Überblick)

- Unterschiede zu traditionellen DBS
 - In Standard-DBS spezifiziert Benutzer Bedingungen, die Ergebnisse erfüllen müssen (bestimmte Attributswerte); deklarative Anfragen in SQL
 - In STMM-DBS sind Anfragen nach bestimmten Attributswerten eher die Ausnahme
 - Typisch: Recherche auf Basis von Ähnlichkeit
 - Spezifikation einer Anfrage durch
 - Konkretes Anfrageobjekt, das durch den Benutzer zur Verfügung gestellt wird (z.B. durch URL, Datei, ...)
 - Vereinfachte Approximation eines Anfrageobjektes (Skizze, Summen, ...)
 - Kein Anfrageobjekt; stattdessen: Datenobjekte werden hierarchisch organisiert, Benutzer navigiert durch diese Hierarchie

– Ähnlichkeitsmodelle

- Modellierung/Repräsentation von Datenobjekten
- Datenspezifisch (Bilder, Sequenzen, ...)
- Anwendungsspezifisch („Personen“, „Abendrot“, ...)

– Ähnlichkeitsmaße

- Messen der Ähnlichkeit zweier Objektrepräsentationen
- Meist werden *Distanzfunktionen* verwendet
- Oft bildet Distanzfunktion D eine *Metrik*

$$- D(x,y) = 0 \Leftrightarrow x = y$$

Reflexivität

$$- D(x,y) \geq 0$$

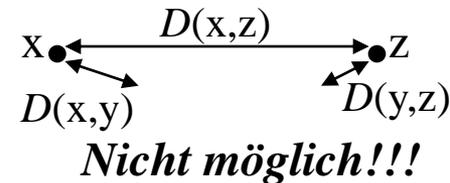
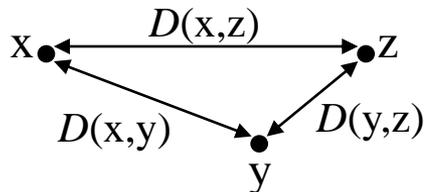
Positiv-Definitheit

$$- D(x,y) = D(y,x)$$

Symmetrie

$$- D(x,y) + D(y,z) \geq D(x,z)$$

Dreiecksungleichung

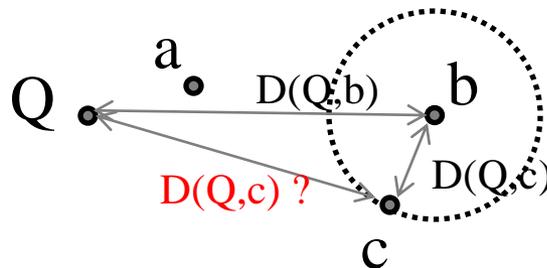


– Dreiecksungleichung sehr wichtig für die Effizienz der Suche

- » Gegeben: $DB=\{a,b,c\}$,
paarweise Distanzen in einer Tabelle
- » Gesucht: ähnlichstes Objekt bzgl. Anfrage Q
- » Ablaufbeispiel:

	a	b	c
a		6.70	7.07
b			2.30
c			

Zuerst wird a mit Abstand 2 von Q gefunden; a ist aktueller Kandidat.
Dann wird b mit Abstand 7.81 von Q gefunden.
Distanz von Q zu c muss nicht mehr berechnet werden, um c auszuschließen, da



$$D(Q,b) \leq D(Q,c) + D(b,c)$$

$$D(Q,b) - D(b,c) \leq D(Q,c)$$

$$7.81 - 2.30 \leq D(Q,c)$$

$$5.51 \leq D(Q,c)$$

- » Objekt c ist also mindestens 5.51 Einheiten von Q entfernt; der aktuelle Kandidat a ist nur 2 Einheiten entfernt

Literatur zur Vorlesung

- Die Vorlesung basiert im Wesentlichen auf aktuellen Forschungsergebnissen
- Meist sind die in dieser Vorlesung besprochenen Konzepte bisher nur in den Originalpublikationen besprochen
- Daher orientiert sich diese Vorlesung leider nicht an einschlägigen Lehrbüchern
- Falls weiterführende Literatur zu einzelnen Aspekten der Vorlesung existiert, wird darauf an entsprechender Stelle hingewiesen.