Skript zur Vorlesung

Neue Trends zur Suche in modernen Datenbanksystemen

Wintersemester 2013/14, LMU München

Vorlesungsteam

Vorlesung (Plenum):

PD Dr. Matthias Renz
Oettingenstr. 67, Zimmer E 1.11
Tel. 089/2180-9331
Sprechstunde: Dienstag, 13^{00}-14^{00}

Übungsbetrieb:

Tobias Emrich
Übungsgruppenleiter
Oettingenstr. 67, Zimmer F 105
Tel. 089/2180-9121

Termine

- Vorlesung: Donnerstag **09:30**–11:45 Uhr, Raum 123, Oettingenstr. 67
- Übung: Mittwoch, 14-16 Uhr Raum M 201 (Hauptgebäude)
  Mittwoch, 16-18 Uhr Raum A 213 (Hauptgebäude)
  (*Start des Übungsbetriebes:* ***30.10.2013***)

Anmeldung für den Übungsbetrieb auf der Homepage

www.dbs.informatik.uni-muenchen.de/cms/Neue_Trends_zur_Suche_in_modernen_Datenbanksystemen

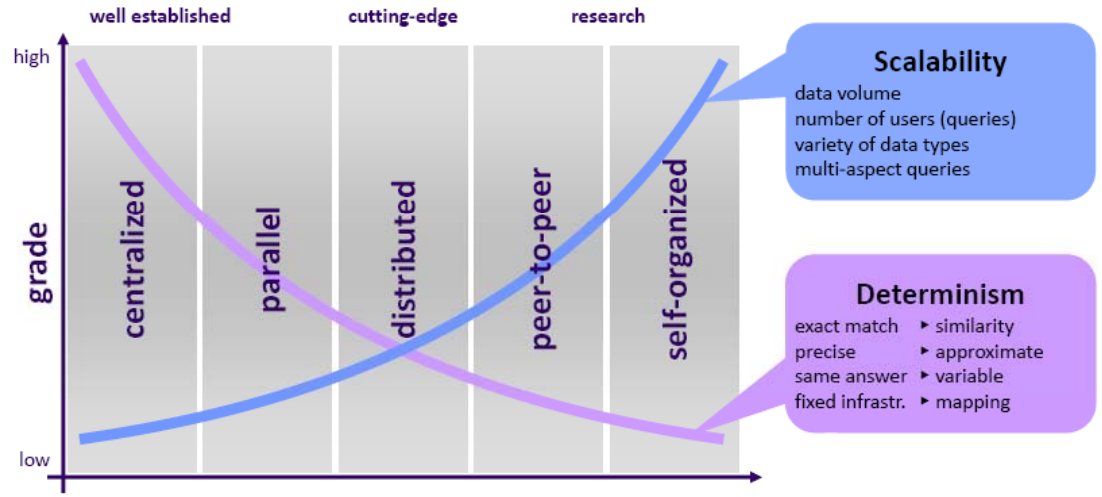
Schein-/Punkteerwerb

- Zulassung: Anmeldung für den Übungsbetrieb (siehe oben)
- Prüfung: Klausur

Kapitel 1
Einführung

Wintersemester 2013/14, LMU München

Trends in Suchstrukturen für moderne Datenbanksysteme



G.Navarro and V.Pestov(Eds.): SISAP2012, LNCS7404, pp.8–24, 2012. © Springer-Verlag Berlin Heidelberg 2012

- Exponentielles Wachstum
  - » des Datenvolumens
  - » der Benutzer / Anfragen
  - » der Vielfalt von Datentypen → Verknüpfung einer Vielzahl von Datenbasen

- Degenerierung der Determiniertheit von Anfrageergebnissen
  - » “Exact match”-Anfragen → Suche nach Ähnlichem
  - » Präzise Anfrageauswertung → Approximative Anfrageauswertung
  - » deterministische Antwort → mehrere/variable Antworten

Inhalt der Vorlesung

- Methoden zur effizienten Bearbeitung von Ähnlichkeits- und Nachbarschaftsanfragen in „modernen“Datenbanksystemen
- Anfragen auf Objekte mit räumlicher, zeitlicher, raumzeitlicher sowie unsicherer Information
- Die in der Vorlesung behandelte Themen:
  - Feature-basierte Ähnlichkeitsanfragebearbeitung
  - Suche nach Objekten mit räumlicher Komponente
  - Suche nach Objekten die sich im Raum bewegen
  - weitere mögliche Ergänzungen
    - Suche nach unsicheren Objekten (Unsicherheit bzgl. der Existenz und/oder bzgl. der Attribute)
    - Suche in Sensornetzwerken, d.h. Suche in Daten die aus einem Sensornetzwerk gewonnen werden

Bezug zur Vorlesung STMD

Stoff dieser Vorlesung ist hauptsächlich aus der Vorlesung STMD II (WS 11/12) übernommen

Themen die ebenfalls in STMD I besprochen wurden sind:

- Grundprinzipien der Feature-basierten Ähnlichkeitssuche
- Basisalgorithmen für Ähnlichkeits- und Nachbarschafts-Anfragen in multidimensionalen Vektorräumen (statische Punktdaten)

Abgrenzung zur Vorlesung DBS I/II

Nicht besprochen werden DBS-Kerngebiete, insb.

- Methoden zur Erfassung und Integrierung der Daten in entsprechende Datenbanksysteme (DBS) / Datenbankmanagementsysteme (DBMS)
- Einbettung und Modellierung der Datentypen nach bestimmten DB-Modellen, wie z.B. das Relationale- / Objekt-Relationale-, sowie Objektorientierte Datenbankmodel
- Umsetzung der Anfragen in gängige Datenbanksystem-Sprachen (DML/DDL) wie SQL

1.1 Gliederung der Vorlesung

1. Einführung
2. Prinzipien der Anfragebearbeitung in modernen DBS
3. Anfragemethoden für Räumlich-Zeitliche Daten
4. Anfragebearbeitung in unsicheren Datenbanken
5. Datenverwaltung und Anfragebearbeitung in Sensornetzwerken

1.2 Warum moderne Datenbanken

- Beschreibung von
  - komplexen Strukturen,
  - (räumliche) Beziehung zwischen Objekten
  - dynamischen Vorgänge
  - Irreversibler Trend in der IT
  - Neue Qualität von Informationen
  - Vermeidung von Informationsverlust durch Integration in Standard-Datenbanksysteme

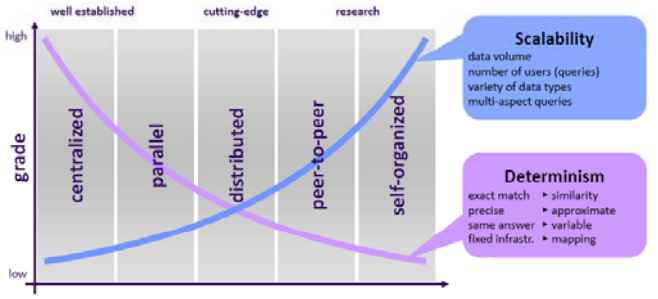


Fig. 1. Development trends in search structures

- Gründe neuer Datenbank-Technologien
  - Sehr große Mengen an Daten vorhanden
  - Speicherplatzintensive Daten
  - In vielen neuen 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

- Was ist ein Modernes 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 räumliche, zeitliche sowie raumzeitliche Datentypen aber auch alphanumerische Datentypen unterstützt
    - das mit großen Datenvolumina umgehen kann

– Überblick: Recherche in modernen DBS

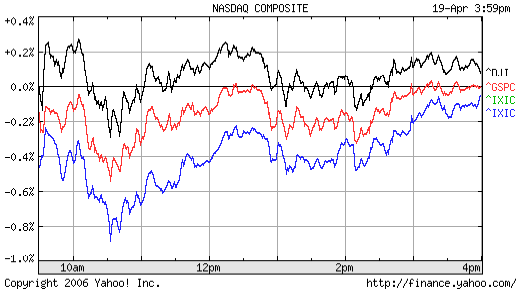
(Unterschiede zur Recherche in traditionellen DBS)

- In Standard-DBS spezifiziert Benutzer Bedingungen, die Ergebnisse erfüllen müssen (bestimmte Attributswerte); deklarative Anfragen in SQL
- In modernen DBS sind Anfragen nach bestimmten Attributswerten eher die Ausnahme
- Typisch: Recherche auf Basis von Ähnlichkeit/Nachbarschaft
- 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, …)

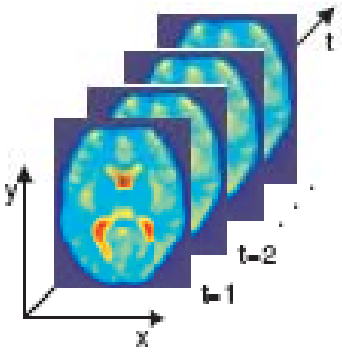
1.3 Anwendungen

1.3.1 Allg. Anwendungen mit räumlichen und raumzeitlichen Daten

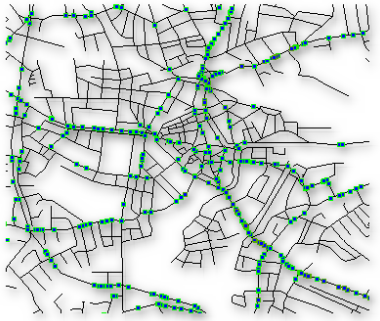
- Prinzipiell alle Anwendungen bei denen Ort- und Zeitinformationen eine Rolle spielen.
- Ort kann auch Position in einem allg. Merkmalsraum (Feature-Raum) sein. → Ähnlichkeitssuche



a) Zeitreihen



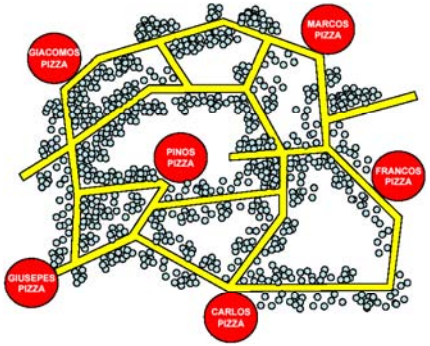
b) CT-Screen-Monitoring



b) Positionen von Autos

1.3.2 Suche nach beweglichen Objekten

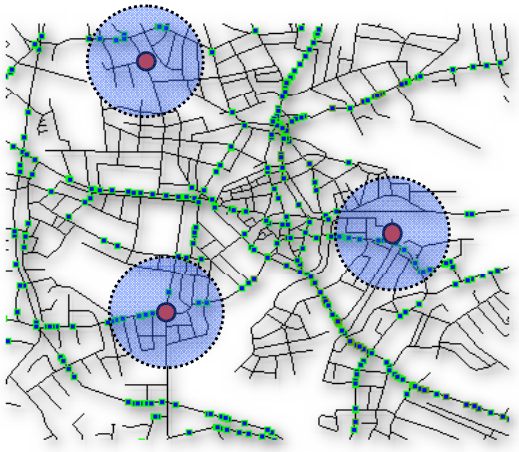
– Analyse von Objekte in einem Verkehrsnetz
  - Location-based Services, Verkehrsplanung und -monitoring
  - Objekte, die sich entlang eines Verkehrsnetzes bewegen
    – Autos, Fußgänger, Züge, …
  - Modellierung
    – Verkehrsnetz ≡ (evtl. gerichteter) Graph
    – Knoten: Kreuzungen, …
    – Kanten: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



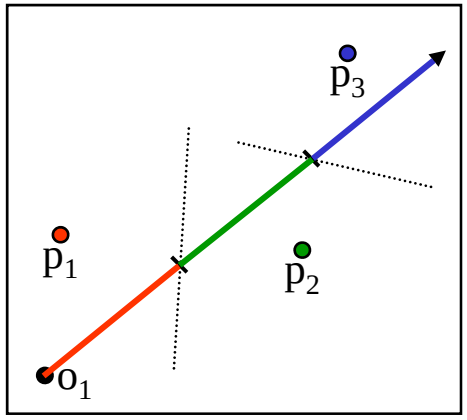
Beispiel: Location-Based Services (LBS)

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



- Objekte die sich in einem (Euklidischen) Raum frei bewegen
  - Raumüberwachung, Chaosforschung, …
  - Modellierung
    - Räumliche Koordinaten
    - Bewegung als Zeitreihe, zu jedem Zeitpunkt Informationen über
      - » Richtung (linear/nicht-linear)
      - » Geschwindigkeit
  - Anfragen/Recherche
    - Suche nach räumlich nahen Objekten
    - Suche nach Objekten mit ähnlichen Bewegungen
    - Suche nach räumlich benachbarten Objekten mit ähnlichem Bewegungsmuster



- Herausforderung
  - Position der Objekte ändert sich ständig

    ➔ Spezielle Methoden zur Konsistenzerhaltung der Datenbank
  - Auswertung des Anfrageprädikates ist teuer: z.B. wird als Distanz bei Straßennetzwerken oft die Netzwerkdistanz verwendet
  - Komplex-strukturierte Anfrageergebnisse: z.B. Anfragen auf unsicherer/nicht exakter Information (z.B. Objekte mit unsicheren Positionen) → Probabilistische Anfragen/Ergebnisse

1.3.3 Suche nach Pfaden in Verkehrsnetzwerken (Navigation)

– Multi-Attributs Pfadsuche in Verkehrsnetzwerken
  - Transportplanung, Navigation, …
  - Gegeben:
    – Transportationsnetzwerk/Strassennetzwerk mit unterschiedlichen Strassenattributen, wie z.B.:
      » Zeit
      » Länge
      » Anzahl von Ampeln
      » etc.
  - Anfragen/Recherche
    – Suche nach optimalen Weg bzgl. mehrerer Attribute

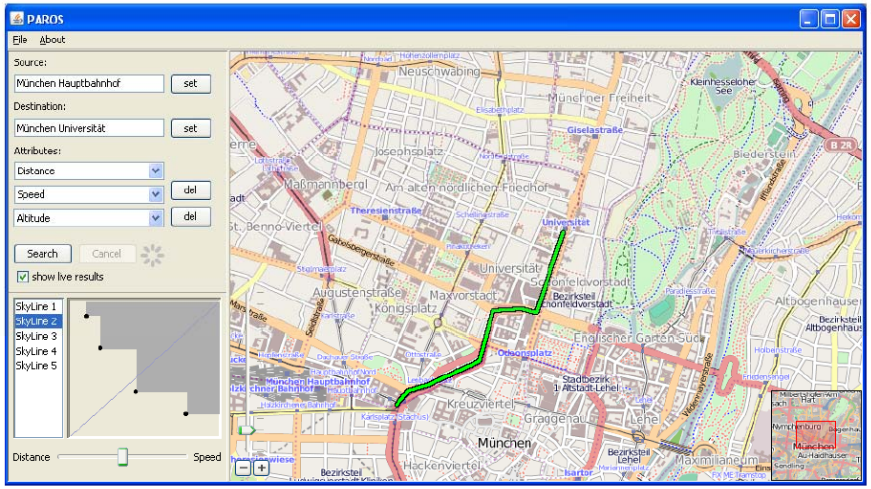


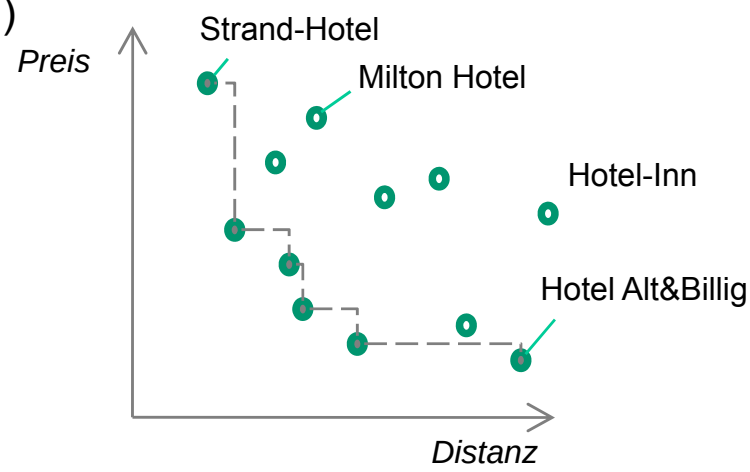
Figure 1: Screen shot of the PAROS interface for 2D Skyline Queries.

– Suche nach Routen mit zusätzlichen Bedingungen
  - Gegeben:
    – Datenbank mit Verkehrsnetzwerkdaten (Straßenkarte, Positionen von interessanten Orten, Geschwindigkeitslimitierungen, etc.)
  - Gesucht:
    – Optimale Route von einem Startpunkt S zu einem Zielpunkt Z, die zuerst an einem Bankautomaten und danach einem Eisladen vorbeigeht.
  - Herausforderungen:
    – Große Suchregion (je nach Abstand zwischen S und T)
    – Berücksichtigung mehrerer Optimierungskriterien (Zeit, Distanz)
    – Limitierter Speicherplatz (Einsatz in „Embedded-Systems“ z.B. Navi)
    – Berücksichtigung von dynamischen Attributen (Attributswerte die sich mit der Zeit ändern, Verkehrslage)

1.3.4 Suche in Empfehlungssystemen (Recommendation Systems)

– Multi-Attributs Ähnlichkeitssuche

- Gegeben:
  - Datenbank mit (Multi-Attribut-) Objekten wie z.B. Produkte (Filme, Elektronische Geräte, Bücher, etc.) oder interessante Standorte (Hotels, Restaurants, Sehenswürdigkeiten, etc.)
- Gesucht:
  - Alle Hotels die sich möglichst in der Nähe vom Strand befinden und möglichst preiswert sind?



- Herausforderungen:
  - Oft negativ-korrelierte Optimierungsvariablen (Attribute)

    (z.B. Preis und Lage)

    => Anstatt eindeutige Antwort wird eine Liste mit unterschiedlichen Empfehlungen zurückgegeben
  - Benutzer haben unterschiedliche Präferenzen:

    1) Unterschiedliche Gewichtungen der Attribute

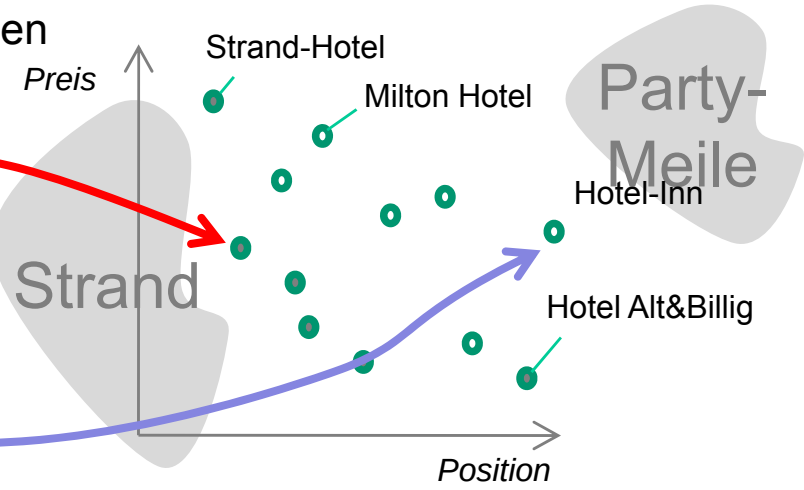
    Fritz Steinreich (B1) „Lieber teurer, dafür näher am Strand“

    Student Kurt (B2) „Für ein paar Schritte mehr, zahle ich gerne nur die Hälfte“

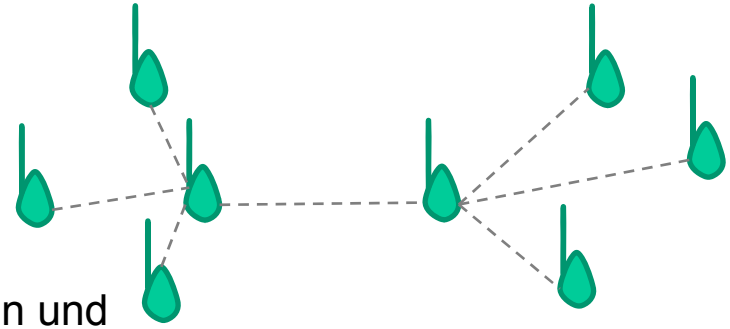
    2) Unterschiedliche Präferenzfunktionen

    z.B.: B1 bevorzugt Distanz zum Strand

    B2 bevorzugt Distanz zur Partymeile



1.3.5 Suche in Sensornetzwerken sensorgestützten Erkundungssystemen (z.B. Environmental Monitoring)



– Gegeben:

  Menge von Sensoren die

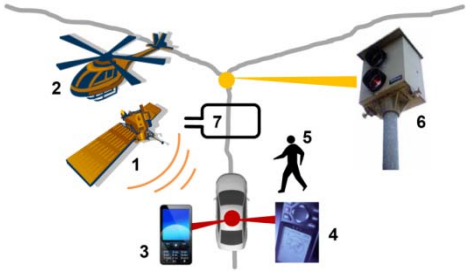
  – jeweils kontinuierlich (oder in regelmäßigen Zeitabständen) Zustände von Prozessen aufnehmen und
  – über ein kabelloses Netzwerk miteinander verbunden sind um miteinander zu kommunizieren und Daten auszutauschen.

  Sensordaten:

  Über die Zeit hinweg Beobachtungen von Ereignissen,

  z.B.:

  – Einfache numerische Attribute wie Temperatur, Luftfeuchtigkeit, etc.
  – Positionen von sich bewegenden Objekten: RFID, Mobile Devices, Radar, etc.



- Weitere Anwendungen:
  - Objektverfolgung
  - Verkehrsanalyse
  - Verkehrskontrollsysteme
- Gesucht:

  Kontinuierliche Ausgabe von
  - Top-k-Anfragen,
  - Aggregationen (Mittelwert, Median, Summe)
  - z.B.: Kontinuierliche Ausgabe (monitoring) der durchschnittlichen Temperatur und Luftfeuchtigkeit
- Ziele und Herausforderungen
  - Minimierung der Übertragungskosten
  - Anfragebearbeitung möglichst in Echtzeit (Realtime)

- Allgemeine Problemklassen (Strategien):
  - Optimierung der Netzwerktopologie (Adhoc-Netzwerke), d.h. Wahl des Weges zur (peer-to-peer) Datenübertragung über das Netzwerk
    - » möglichst kurze Übertragungswege
    - » möglichst wenig Einzel-Übertragungen
  - In-Network Anfragen:
    - » Teile der Anfrage werden bereits in den Netzwerkknoten verarbeitet
  - Extraktion und Weiterleitung von relevanter Information (Filtering)
- Herausforderungen:
  - Hoch dynamische Daten
  - Sehr große Datenfluten
  - Resourcen-limitierte Sensorgeräte

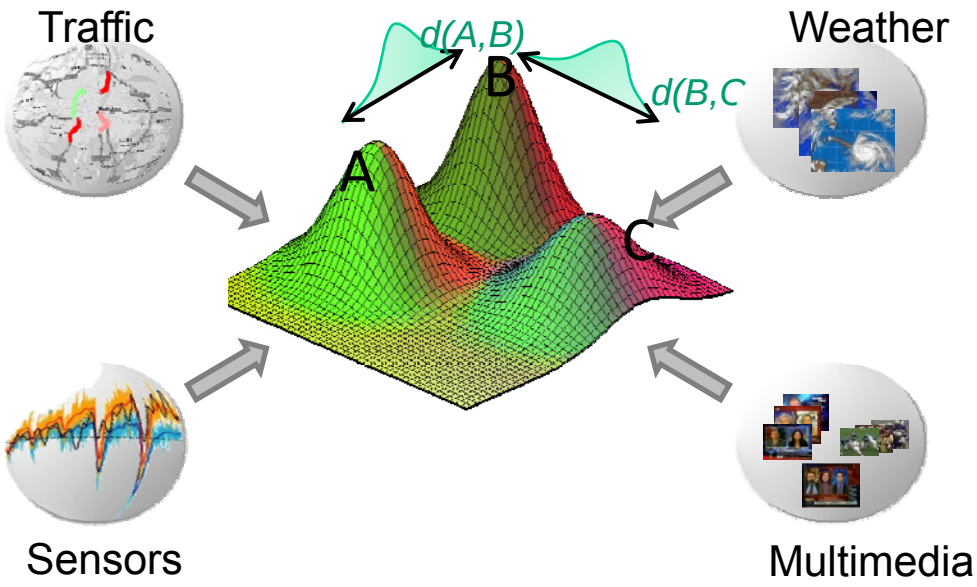
    → wenig Speicherplatz

    → Minimierung der CPU Kosten

    → Minimierung der Kommunikationskosten

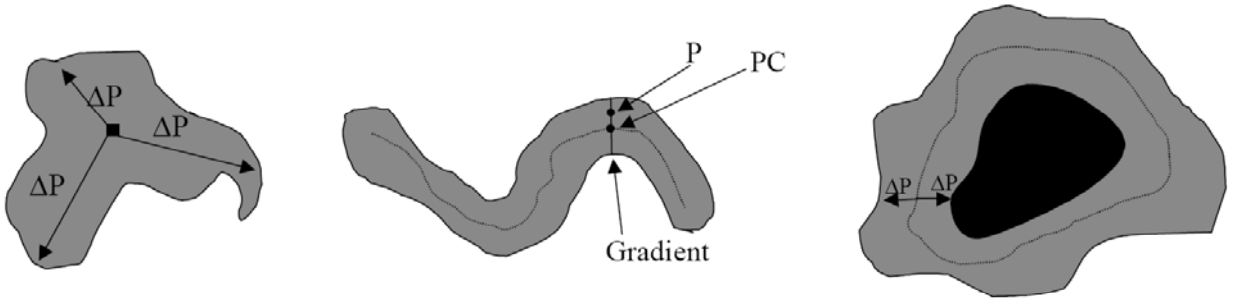
1.3.6 Suche in unsicheren Datenbeständen

– Wahrscheinlichkeitsbasierte Anfragen auf räumlich unsicheren Objekten



- Unsicherheit der Daten durch
  - Ungenauigkeit der Positionssensoren (GPS, Radar, etc.)
  - Diskrete Abtastung (Snapshots) von sich zeitlich kontinuierlich verändernden Variablen (RFID-tracking Systeme)
  - Vorhersagen (Aktienkurse, Orkane, Positionen von Eisbergen, ..)
  - Maßnahmen zur Bewahrung der Datenschutzrichtlinien (Depersonalisation)

- Beispiele von räumlich unsicheren Objekten



a) unsichere Position b) unsicherer Pfad c) unsichere Region

- Modellierung

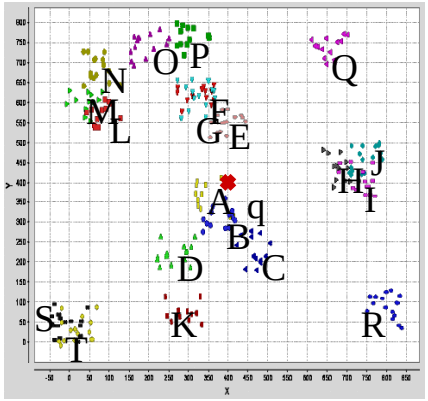
abhängig vom Objekttyp (siehe Skizze oben)

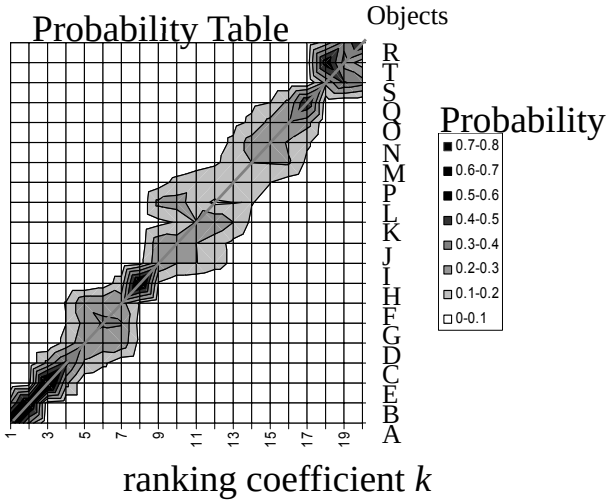
– Punkt (→ Verteilungsfunktion)
– Trajektorie (→ Erwartete Trajektorie mit Varianzangaben)
– Region (→ Wahrscheinlichkeit über die Objektzugehörigkeit von Punkten („fuzzy objects“))

abhängig von der Art der Unsicherheit

– existentielle Unsicherheit (Tupel-Unsicherheit) (→ Tupel + Wahrscheinlichkeit)
– Attributunsicherheit (Unsicherheit der Ausprägung der Attribute (z.B. Ort))

- Anfragen/Recherche
  - Probabilistische Ähnlichkeits-/Nachbarschaftsanfragen
  - Wahrscheinlichkeitsverteilung über Verkehrsdichte
  - Probabilistische Räumliche Anfragen (z.B. Schnittanfragen)
  - Probabilistische Raum-Zeit-Anfragen
  - etc.

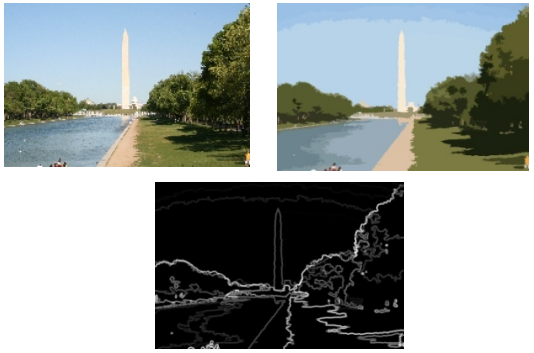




a) Unsichere Punktobjekte b) Probabilistisches Distanzranking-Ergebnis

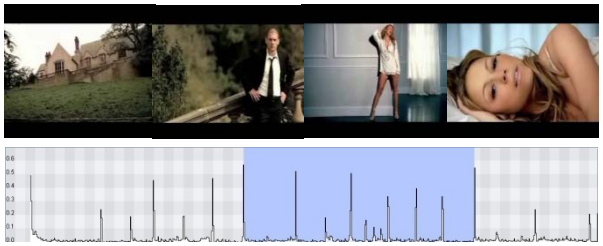
1.4.5 Suche in Multimediadaten

– Suche in Multimediadaten



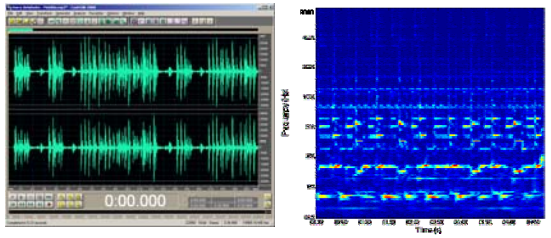
a) Bilddaten

- Ähnlichkeitssuche
- Objekterkennung
- Content-Based Image Retrieval
- ...



b) Videodaten

- Shot Detection
- Motion Tracking
- Video Similarity
- Video Indexing
- Objekterkennung
- ...



c) Audiodaten

- Title Identification
- Genre Klassifikation
- Query by Humming
- Speech Recognition
- ...

– Allg. Anfragemethodik: merkmalsbasierte Ähnlichkeitssuche
– Spezielle Methoden zur Merkmalsextraktion sind nicht Fokus dieser Vorlesung

1.4 Allg. Problematiken bei der Suche in großen Datenbeständen

- 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)
    → **Indexierung**
  - Beschleunigung der Einzelvergleiche (geeignete Repräsentationen)
    → **Mehrstufige Anfragebearbeitung**

1.5 Literatur zur Vorlesung

- Die Vorlesung basiert im Wesentlichen auf aktuellen Forschungsergebnissen
- Meist sind die in dieser Vorlesung besprochenen Konzepte bisher nur in den Orginalpublikationen 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.