C | A | U Christian-Albrechts-Universität zu Kiel Technische Fakultät

Skript zur Vorlesung infEAeS-01a - Methoden der Effizienten Ähnlichkeitssuche in großen Datenbeständen

Wintersemester 20/21, CAU Kiel

Dozent: Prof. Dr. Matthias Renz

Übung: Christian Beth / Niko Amann

infEAeS-01a - Methoden der Effizienten Ähnlichkeitssuche in großen Datenbeständen

Wintersemester 20/21, CAU Kiel

Kapitel 1: Einführung

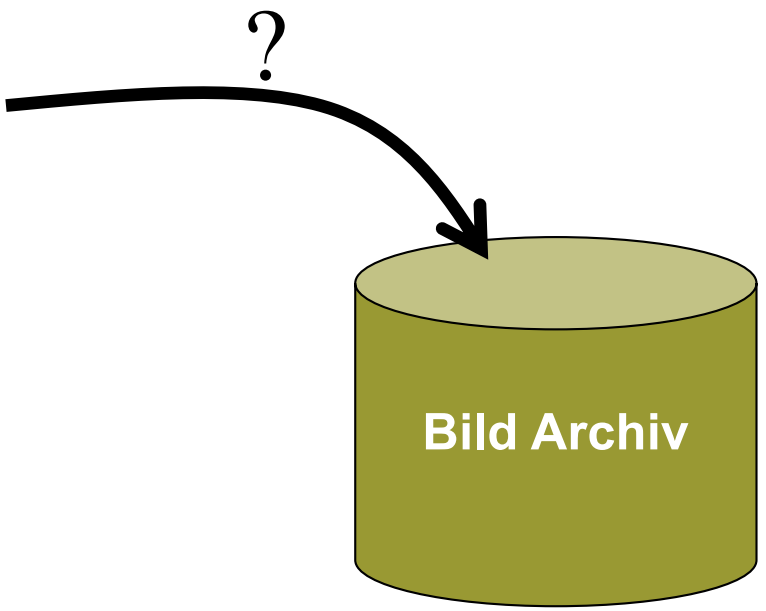
1 Einführung

- Inhalt der Vorlesung
  - Algorithmen und Indexstrukturen für Ähnlichkeits- und Nachbarschafts-anfragen in großen Datensammlungen mit komplex-strukturierten Objekten:
    - aus der Wissenschaft (Medizin, Biologie, Archäologie, …)
    - Multimedia Objekte (Bilder, Videos, …)
    - Räumlich ausgedehnte Objekte (CAD Daten)
    - Zeitliche (zeitabhängigen) Objekte bzw. Sequenzdaten (Temperaturkurven, Aktienkurse, …)
  - Merkmalsbasierte (Feature-basierte) Ähnlichkeitsmodelle

    Merkmalsextraktion aus räumlichen, zeitlichen und Multimedia Objekten

1.1 Motivierende Beispiele

- Beispiel 1: Bildersuche
  - Gegeben: Archiv mit 2 Mio. Bildern (2D Objekte)
  - Frage: Ist im Archiv ein bestimmtes Kunstwerk abgebildet?



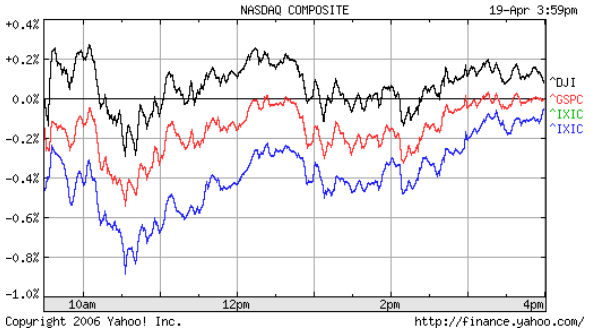
1.1 Motivierende Beispiele

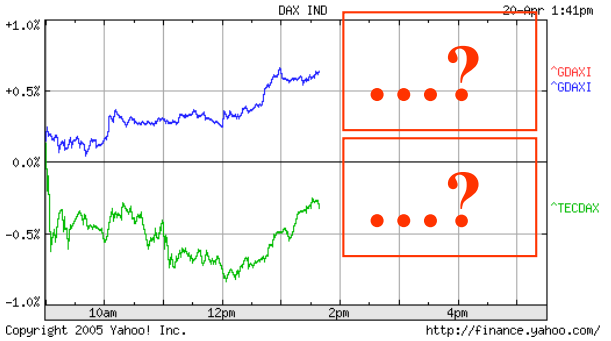
- Herausforderung
  - „abgebildet“ heißt nicht „identische Binärrepräsentation“ wie das Anfragebild

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

1.1 Motivierende Beispiele

- *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
    - …





1.1 Motivierende Beispiele

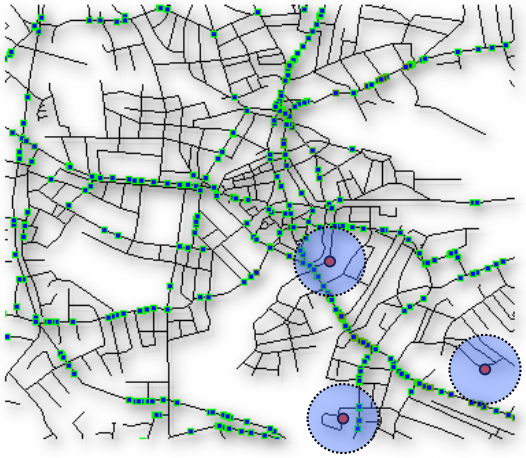
- *Beispiel 3: Sequenzsuche in Videodateien*
  - Gegeben: Datenbank von Videofilmen, Anfragesequenz
  - Gesucht: alle Videos in der DB, die eine Sequenz ähnlich der Anfragesequenz enthalten
  - Herausforderungen:
    - Erkennen von Bildinhalten statt reiner Bildähnlichkeit
    - Unvollständige Bildsequenzen
    - Unterschiedlich lange Sequenzen
    - Unterschiedliche „Auflösung“ (FPS, Bildauflösung)
    - …

1.1 Motivierende Beispiele

- *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
    - …



1.2 Probleme der Ähnlichkeitssuche - Einführung

- 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“)

1.2 Probleme der Ähnlichkeitssuche - Einführung

- Allgemeine Problemstellungen bei der Ähnlichkeitssuche (cont.)
  - Pragmatische Ebene
    - (effizienter) Algorithmus zur Bestimmung der Ähnlichkeit zwischen zwei Objekten
    - (effizienter) Algorithmus zur Suche von ähnlichen Objekten in einer großen Datenbank

1.2 Probleme der Ähnlichkeitssuche - Einführung

- 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)

1.2 Probleme der Ähnlichkeitssuche - Einführung

- 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
  - Somit gelten die folgenden beiden Bedingungen:

    Bed. 1: s(v) = s(w) => (w stellt v dar),
    d.h. (w stellt nicht v dar) => s(v) ≠ s(w)
    Bed. 2: (w stellt v dar) => s(v) = s(w),
    d.h. s(v) ≠ s(w) => (w stellt nicht v dar)
  - Bewährte Suchtechniken (standardmäßig integriert in standard DBMS) skalieren gut für sehr große Datenbanken (Suchbaum, Hashverfahren)
  - ABER: geeignete Normalform(en) aufgrund von Bed. 1 (s.o.) schwierig (sehr unwahrscheinlich) zu finden.

1.2 Probleme der Ähnlichkeitssuche - Einführung

- Lösungsansatz 2: Feature-basierte Ähnlichkeit
  - Beispiel:
    Bildsuche: Einfache Eigenschaft eines Bildes: Durchschnittsfarbe

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

    dann gilt

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

    und somit gilt $\text{avg}(v) \neq \text{avg}(w) \Rightarrow$ (w stellt nicht v dar)

    Nach diesem Prinzip können folgende Ähnlichkeitsanfragen spezifiziert werden:

    wenn $\mathbf{|avg(v) - avg(w)| \leq \varepsilon}$, dann ist **v ähnlich zu w**

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

1.2 Probleme der Ähnlichkeitssuche - Einführung

- Lösungsansatz 2: Feature-basierte Ähnlichkeit (cont.)
  - Mögliche Erweiterungen
    - Farbhistogramme statt einfache Durchschnittsfarbe
    - Beziehungen (Korrelationen) der einzelnen Dimensionen berücksichtigen
    - Berücksichtigung der dargestellten Formen (geometrische Ebene)
    - Berücksichtigung von dargestellten Objekten (semantische Ebene)
    - …