Kapitel 1 Einführung

1.1 Beispiel: Bildsuche

Gegeben: Archiv mit 100,000 Bildern (aus dem WWW, aus einem Katalog, …)

Frage: Ist im Archiv ein bestimmtes Kunstwerk abgebildet (als ganzes)?

Herausforderung:

- "abgebildet" bedeutet nicht „identische Binärrepresentation" wie das Musterbild

- Abweichungen im Beispiel „Sonnenblumen"
  - — unterschiedliche Größe (Skalierung; Auflösung)
  - — unterschiedliche Tönung der Farben
  - — abweichende Ausschnittbildung
  - — hinzugefügter Rand oder Beschriftung
  - — …

Die Aufgabe führt zu folgenden Problemstellungen:

- Informelle Ebene
  - Ähnlichkeit kann situationsabhängig sein, z.B.:
    - bei Suche nach „Abendrot“: Farben wichtig
    - bei Suche nach „Personen“: Formen wichtig
  - Ähnlichkeit kann personenabhängig sein (z.B. rot-/grün-Blindheit)
  - Erkennung von Ähnlichkeit ist Gegenstand von psychologischer und perzeptionswissenschaftlicher Forschung

- Formale Ebene
  - mathematische Beschreibung von Bildern
  - mathematische Beschreibung von „Ähnlichkeit“ als Vergleich von Bildern
  - Ähnlichkeitsmaß: Bewertung der Ähnlichkeit zweier Bilder durch eine Maßzahl (Bsp. „100% ähnlich“ oder komplementär: „Abstand gleich 0“)

- Pragmatische Ebene
  - (effizienter) Algorithmus zur Berechnung der Ähnlichkeit zweier Bilder
  - (effizienter) Algorithmus zur Suche von ähnlichen Bildern in einer Datenbank

Teilproblem der Suche

- Sequentielle Suche
  - Ablauf: Vergleich des Anfragebildes mit jedem einzelnen Bild der Datenbank.
  - dauert „viel zu lange“, d.h. skaliert *linear* zur Größe der Datenbank:
    100-mal mehr Bilder → 100-mal längere Suchzeit.
- Herausforderung
  - Beschleunigung der Suche durch geschickte Datenorganisation (Indexstrukturen)
  - Schnellere Einzelvergleiche durch geeignete Repräsentation (z.B. Approximation)

Lösungsansatz 1: Annahme einer Normalform

- *Normalform:* Es gibt eine Stringdarstellung s(v), s(w) für jedes Bild v, w, sodass gilt:

  $$s(v) = s(w) \quad \Leftrightarrow \quad w \text{ stellt } v \text{ dar}$$

- geeignete Normalform(en) zu finden ist sehr schwierig / sehr unwahrscheinlich
- Suchtechniken sind bewährt und skalieren sehr gut für sehr große Datenbanken
  - Suchbaum $O(\log n)$
  - Hashverfahren $O(1)$

Lösungsansatz 2: Featuretransformation

- Beispiel: Durchschnittsfarbe eines Bildes

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

  dann gilt:

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

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

  — sinnvoll, falls nicht zu viele Bilder ε-ähnlich sind.
  — mehrstufiges Vorgehen: avg als Filter, genauer Vergleich als Verfeinerung

- naheliegende Erweiterungen der Idee
  — Farbhistogramme statt einfacher Durchschnittsfarbe
  — Anpassung der Dimensionen

- weiterführende Ansätze
  — Berücksichtigung von dargestellten Formen (geometrische Ebene)
  — Berücksichtigung von dargestellten Objekten (semantische Ebene)
  — Erweiterung auf Bildfolgen (Videos)

1.2 Inhalt der Vorlesung

- Medien und Multimedia-Datenbanken
  — Kurze Einführung
  — Schwerpunkt der Vorlesung liegt auf Recherche / Retrieval in MM-Datenbanken

- Modellierung von Ähnlichkeit
  — Ziel: Mathematische Modellierung von Ähnlichkeit an realen Beispielen.
  — Aufgreifen verschiedener Anwendungen: Sequenzdaten/Zeitreihen, Farben und Formen in Bildern, geometrische Objekte, strukturierte Objekte (Graphen, Bäume)

- Algorithmen zur Ähnlichkeitssuche
  — Ziel: effiziente Bearbeitung von Ähnlichkeitsanfragen auf sehr großen Datenbanken
  — Mehrstufige Anfragebearbeitung: Filter- und Verfeinerungstechniken
  — Approximationstechniken zum schnellen eins-zu-eins-Vergleich
  — Approximationen, die sich zur Indexbildung eignen
  — Indexstrukturen zur schnellen Generierung von Kandidatenmengen
  — Analysen und Kostenmodelle zur Vorhersage der Bearbeitungszeiten von Anfragen bzw. zur dynamischen Auswahl verschiedener möglicher Bearbeitungswege

1.3 Warum Multimedia-Datenbanken?

Multimedia

- Multimedia ist ein irreversibler Trend in der Informationstechnologie
- Multimedia verbessert die Qualität von Information
- geringerer Informationsverlust, wenn die Ein- und Ausgabe im jeweiligen Medium direkt (ohne Umsetzung in ein anderes Medium, z.B. Text) geschieht

Warum Datenbank-Technologie für Multimedia?

- Multimediale Information ist speicherplatzintensiv
- Multimediale Information soll für viele Benutzer gleichzeitig zugreifbar sein und soll auf konsistente Weise bearbeitet und manipuliert werden können
- Multimediale Information soll recherchierbar sein

Standard-Datenbanksysteme

erlauben es, daß viele Benutzer auf große Mengen an (relationalen) Informationen gleichzeitig über LAN oder WAN zugreifen und diese unter Konsistenzerhaltung bearbeiten.

Wichtige Eigenschaften eines Datenbank-Managementsystems beim gleichzeitigen Zugriff auf gemeinsam genutzte Informationen sind:

- Physische und logische Datenunabhängigkeit
- Anfragebearbeitung d.h. Unterstützung der Recherche in sehr großen Datenbeständen
- Speicherungsstrukturen für den effizienten Zugriff
- Unterstützung von Transaktionen
  - Concurrency: Gleichzeitige Updates mehrerer Benutzer werden voneinander isoliert
  - Recovery: Konsistentes Wiederaufsetzen im Fehlerfall
  - Die Integrität des Datenbestandes wird überwacht
- Daten-Sicherheit und Datenschutz.

Was versteht man unter einem Multimedia-Datenbanksystem

Der Begriff “Multimedia-DBS” wird verschieden verwendet, z.B. für:

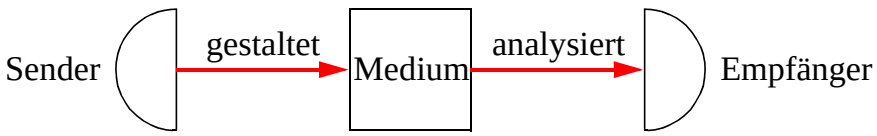
- CD-ROM-Sammlungen, die umfassende Information unter einem Stichwort zugreifbar machen
- Systeme, die dem Benutzer durch Hilfsmittel helfen, multimediale Information zu organisieren und zu sichten (Browser)
- Video-on-demand-Systeme
- CAD-Systeme, die eine Datenbank als Speichersystem nutzen
- Relationale Datenbanksysteme, die zusätzlich zur tabellarischen Information sog. “Binary Large OBjects” (BLOB) speichern

Wir werden unter einem Multimedia-Datenbanksystem ein Datenbanksystem mit hoher Kapazität und hoher Performanz verstehen, das sowohl Multimedia-Datentypen als auch alphanumerische Datentypen unterstützt, und das mit großen Volumina von (insbes. multimedialer) Information umgehen kann.

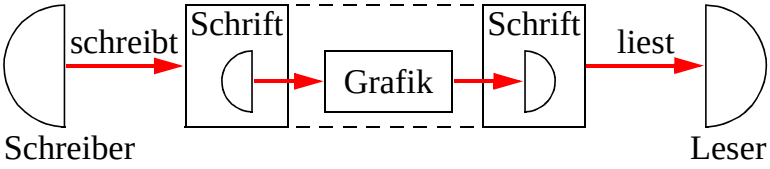
1.4 Was ist ein Medium?

- Laut Duden: Mittelglied, Mittler, Mittelsperson
- Umgangssprachlich: Presse, Rundfunk, Fernsehen, Internet
- Physikalisch: Schallwellen, elektromagnetische Wellen
- Hardwaretechnisch: Speichermedium, Übertragungsmedium (Telefonleitung, Funk)
- Logisch: Abstrakter Datentyp (Text, Grafik, Videodatei usw.)

Allen diesen Definitionen ist gemeinsam, daß Medien *Nachrichten übermitteln.* Das Medium “vermittelt” zwischen zwei oder mehr Kommunikationspartnern. Der *Sender* einer Nachricht erzeugt *Signale* (Verlauf einer physikalischen Größe) im Medium. Der *Empfänger* (*Rezeptor*) nimmt das Signal auf, um es zu interpretieren, zu wandeln und weiter zu leiten. Das Medium ist ein *Nachrichtenträger*:

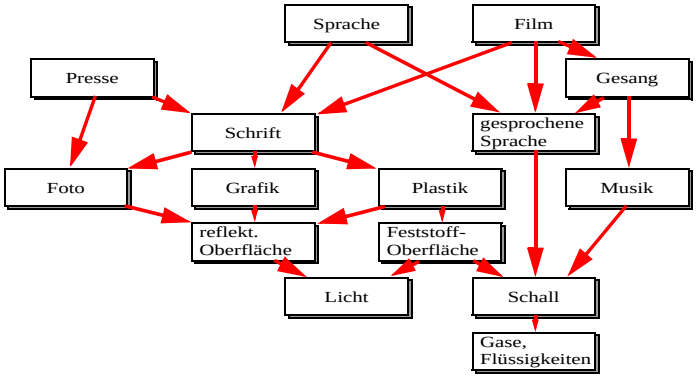


Unter den unterschiedlichen Medien gibt es eine Benutzungshierarchie. So bedient sich das Medium Text (Schrift) z.B. grafischer Symbole (Buchstaben), um die Information darzustellen. Das Medium *Schrift* benutzt also das Medium *Grafik*.



- Manche Medien benutzen auch mehrere Medien (z.B. *Presse* benutzt *Schrift* und *Bild*)

Benutzungshierarchie einiger Medien (nach [Meyer-Wegener 1991]):



Einteilung der Medien nach menschl. Sinnesorganen [Meyer-Wegener 1991]:

- Medien
  - visuell (optisch)
    - sprachlich (symbolgebunden)
      - Text
      - Piktogramme
    - nicht-sprachlich
      - Grafik
      - Foto
  - akustisch
    - Sprache
    - Musik
    - Geräusche
  - haptisch (tasten)
  - gustorisch (schmecken)
  - olfaktorisch (riechen)

1.5 Typen von Multimedia-Daten

Text

Dieser Medientyp ist oft auf die Repräsentation von Strings von Zeichen reduziert. Eine sinnvolle Repräsentation in einem Multimedia-Dokumentenarchiv sollte zusätzlich folgende Möglichkeiten vorsehen:

- Strukturelle Information wie z.B. Titel, Autor, Kapitel, Abschnitt usw. wie dies durch sog. Markup-Languages (SGML, HTML usw.) möglich ist.
- Layout-Information

Selbst mit Formatierungsinformation ist Text das am wenigsten speicherplatzintensive Medium. Ohne Kompression belegt eine A4-Seite etwa 2KBytes. Benutzer erwarten bei Text in Multimedia-Datenbanken im allgemeinen gute Recherchemöglichkeiten nach einzelnen Worten und Wortkombinationen (Volltextsuche), wobei fortgeschrittene Systeme Features bieten wie z.B. Toleranz gegenüber Orthographie- oder Konvertierungsfehlern (OCR), Synonyme, sowie das Suchen ähnlicher Textdokumente.

Rasterbild (still image)

Entstehen meist durch Eingabe von einer digitalen Kamera oder einem Scanner, können aber auch aus anderen Daten (Texte, Grafiken, Meßwerte) erzeugt werden. Es gibt unzählige verschiedene Formate und Kompressionsverfahren (GIF, TIFF, JPEG, ...).

Ein Rasterbild ist konzeptuell eine Matrix von Bildpunkten (Pixel). Zur Darstellung eines Pixel werden unterschiedlich viele Bits verwendet:

- **Bilder mit 2 Farben**

  1 Bit (schwarz/weiß, bzw. Vordergrund-/Hintergrundfarbe)

  Ausschnitts- und Überlagerungsfunktionen sehr effizient, weil lediglich bitweise boolsche Operatoren (UND, ODER) angewandt werden.

- **Grauwert- und Farbbilder:**

  benötigen mehr als 1 Bit pro Pixel. Die Anzahl der Bit pro Pixel (z.B. 24) wird auch als *Farbtiefe* oder *Pixeltiefe* bezeichnet. Die Pixeltiefe ist innerhalb eines Rasterbildes immer konstant. Bei verschiedenen Bildern in einer Bilddatenbank kann die Farbtiefe aber variieren. Sie muß deshalb in der Datenbank gespeichert werden. Stehen nur wenige Bit pro Pixel (z.B. 8) zur Verfügung, dann werden meist Farbtabellen angelegt. Die Farbtabelle ordnet jeder Bitkombination einen Farbwert zu, der meist im RGB-Farbmodell codiert wird: Je 8 Bit für die drei Grundfarben *rot*, *grün* und *blau*, aus denen

sich durch Mischung alle Farben erzeugen lassen. Alternativ gibt es auch Farbmodelle mit anderen Grundfarben (CYM) oder mit Farbwert, Helligkeit und Sättigung (YIQ). Bei höherer Pixeltiefe werden die Farbwerte meist direkt (ohne die Umsetzung durch eine Farbtabelle) in den Pixel gespeichert, da die Farbtabelle zu groß würde.

Höhe und *Breite* eines Bildes müssen ebenfalls bekannt sein, um das Bild korrekt darzustellen (es gibt i.A. kein spezielles “Zeilenende-Zeichen” in Rasterbildern).

Der Speicherbedarf für ein Rasterbild variiert je nach Anwendung und Kompressionsverfahren sehr stark. Er reicht von einigen KBytes für komprimierte schwarz/weiß-Bilder bis zu mehreren 100 MBytes für Satelliten-Aufnahmen.

Basisoperationen, die ein Multimedia-DBMS unterstützen sollte, sind etwa die Selektion von Teilbildern oder das Skalieren (Ändern der Auflösung). Es gibt zahlreiche verschiedene Ansätze zur Recherche in Bilddatenbanken, die v.a. in Kapitel 2 der Vorlesung behandelt werden. Sie reichen von manuell erfaßten textuellen Beschreibungen bis zur vollautomatischen Suche nach Bildern mit ähnlichem Inhalt (d.h. ähnlichen Farben, Formen, Texturen usw.)

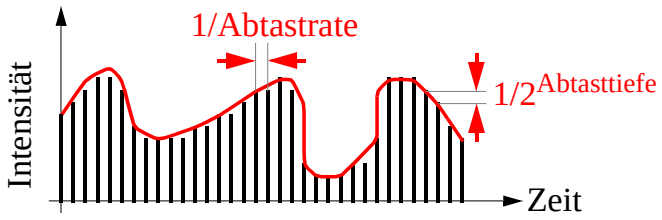
Grafik (Vektorgrafik)

Im Gegensatz zu Rasterbildern enthalten Objekte des Mediums Grafik eine abstraktere Beschreibung des visuell dargestellten Inhalts. Eine Grafik ist meist hierarchisch aus Grundformen wie Linie, Kreis, Polygon zusammengesetzt. Der wichtigste Vorteil gegenüber Rasterbildern ist die bessere Manipulierbarkeit durch den Benutzer. Die Grundformen lassen sich einzeln selektieren und bearbeiten. Bei Rastergrafiken ist die Selektion einzelner Teilobjekte schwierig oder unmöglich. Vektorgrafiken sind je nach Komplexität typischerweise speichersparender als Rasterbilder.

Wegen der fehlenden Normung der Grafikformate gibt es jedoch kaum Ansätze, die Manipulations- und Rechercheoperatoren in Datenbanksysteme integrieren.

Audio (Musik, Geräusch, Sprache)

Durch die breite Verfügbarkeit von leistungsfähigen Soundkarten und CD-ROM-Laufwerken ist die Verwendung des Mediums Audio sehr populär geworden. Die gesprochene Sprache ist das mit Abstand wichtigste Kommunikationsmedium des Menschen.



Die einzelnen Medien Musik, Geräusch und Sprache unterscheiden sich in erster Linie im Hinblick auf die Qualitätsanforderungen, aber auch in Bezug auf die Rechercheanforderungen. Die einfachste Darstellungsform einer Audiosequenz stellt die sog. *Pulse Code Modulation* (*PCM*) dar, die das Signal in festen Zeitabständen (gemäß der *Abtastrate, sampling rate*) abliest und den Meßwert dann mit einer festen Anzahl von Bits (der *Abtasttiefe*) codiert:

Die Abtastrate und die Abtasttiefe bestimmen die Qualität der Audiosequenz. Gemäß dem Abtasttheorem muß die Abtastrate mindestens doppelt so hoch sein wie die höchste vorkommende Frequenz (Grenzfrequenz) des Audiosignals. Folgende Tabelle zeigt typische Qualitätsstufen:

Qualitätsstufe	Grenzfrequenz	Abtastrate
Telefon	3000 Hz	6000 Samp./s
Mittelwellenradio	4000 Hz	8000 Samp./s
UKW-Radio	8000 Hz	16000 Samp./s
Hifi (CD)	22000 Hz	44000 Samp./s

Für CD-Qualität wird jeder der beiden Stereokanäle (*links/rechts*) mit einer Abtasttiefe von jeweils 16 Bit abgetastet. Daraus ergibt sich ein Speicherbedarf für 75 Minuten (max. Kapazität einer CD) von 2·16·44000·60·75/8 Bytes = 755 MBytes (also 10 MBytes pro Minute). Es gibt zahlreiche Kompressionsverfahren für Audio. Der minimale Speicherplatzbedarf für Audio in Telefonqualität liegt bei 20 KBytes pro Minute.

Wichtige Operationen, die das DBMS unterstützen sollte, sind das Selektieren von Teilen einer Tonaufnahmen gem. Start- und Endzeitpunkt und das Aneinanderfügen und Mischen von Audiosequenzen, also typische “Schneideoperationen”, Operationen zur Beeinflussung der Lautstärke und der Qualität (Abtastrate, Abtasttiefe).

In Analogie zu Texten und Bildern kann die Recherche in großen Mengen von Audiosequenzen über Mustererkennung erfolgen. Ähnliche Muster sollen meist ohne Berücksichtigung der Lautstärke und Sprechgeschwindigkeit gefunden werden, sowie Sprecher-unabhängig.

Eine Sonderstellung unter den Audiomedien nimmt Musik ein, die nicht als Folge von Abtastwerten sondern in abstrakter Form (z.B. als Notensequenz oder MIDI-Daten) dargestellt wird, um mit einem Synthesizer abgespielt zu werden. Das Verhältnis zwischen diesen beiden Medien ist ähnlich wie das von Rasterbild zu Vektorgrafik.

Video

Wir verstehen unter Video eine Aggregation von Rasterbildern, die in einem strengen zeitlichen Bezug zueinander stehen. Um einen Eindruck von einer einigermaßen kontinuierlichen Bewegung zu bekommen, müssen mindestens 25 Einzelbilder pro Sekunde abgespeichert werden. Es gibt auch Spezialanwendungen (Animation) bei denen statt Rasterbildern Vektorgrafiken verwendet werden. Oft besteht bei Video die Anforderung der Synchronisation mit weiteren Medien (Audio, Text).

Video stellt die höchsten Anforderungen an die Speicherkapazität. Bei 25 Bildern pro Sekunde und einem Speicherbedarf von 250 KByte pro Einzelbild ergibt sich ein Speicherbedarf von 375 MByte pro Minute. Dies macht Kompressionsverfahren unumgänglich.

DBMS sollten die üblichen Schneide-Operationen für Video unterstützen, sowie die Selektion von Standbildern und die Konvertierung in bestimmte Datenformate, Qualitätsstufen und Kompressionsverfahren. Rechercheverfahren erfordern die Zerlegung der Videos in Szenen, Einstellungen usw.

CAD-Bauteile

Wir wollen im Rahmen dieser Vorlesung auch einige Medien behandeln, die nicht im landläufigen Sinn als Medien verstanden werden. CAD-Bauteile werden je nach Modellierungstechnik als dreidimensionale Erweiterung von Rasterbildern oder Vektorgrafiken verstanden und werden auch in ähnlicher Weise wie andere Medienobjekte in typischen Multimedia-Anwendungen (z.B. Schulung, Werbung) verwendet und bearbeitet.

Auch die Recherche in großen Datenbeständen ist bei diesem Medium von zentraler Bedeutung. Man denke beispielsweise an das Patentamt oder an Automobilhersteller, die eine große Vielfalt an Bauteilen und Varianten zu verwalten haben. Bei der Recherche steht die geometrische Form des Bauteils im Vordergrund des Interesses.

Moleküle

Auch Moleküle sind kein Medium im landläufigen Sinn, werden aber im Unterricht, in der Werbung und anderen Multimedia-Anwendungen verwendet. Bei der Recherche sind die Geometrie sowie die chemischen Eigenschaften v.a. der Molekül-*Oberfläche* von zentraler Bedeutung. Wichtig ist nicht nur die Suche nach Objekten mit ähnlicher Geometrie sondern auch die Suche nach Objekten mit komplementärer Geometrie, da solche Moleküle Interaktionspartner sind, was für die pharmazeutische Forschung wichtig ist.

Graphen

Viele Konzepte des menschlichen Denkens werden durch Graphen beschrieben, z.B. elektrische Schaltpläne, Organisationsstrukturen usw. Damit wird auch der Graph zu einem Medium, das von einem Multimedia-Datenbanksystem unterstützt werden soll. Recherchemechanismen sollten sowohl die Struktur des Graphen als auch Knoten- und/oder Kantenbeschriftungen (oder -Bewertungen) umfassen und sowohl Gesamtähnlichkeit (ganzer Graph ähnlich) als auch partielle Ähnlichkeit (Teilgraph ähnlich) ermitteln.

Zeitreihen

Zeitreihen sind meist Meßwerte einer physikalischen Größe, die periodisch aufgenommen werden. Prinzipiell ist auch das Medium Audio eine Zeitreihe. Es wurde hier aber extra angesprochen, weil einige Unterschiede bestehen: Bei Audio steht die Zeitgenaue Wiedergabe des Mediums im Vordergrund, die bei typischen Zeitreihen wie z.B. Aktienkursen, Fieberkurven usw. keine Rolle spielt. Wichtig ist für die Recherche das Auffinden von ähnlichen Formationen in den Zeitreihen.

1.6 Recherche in Multimedia-Datenbanken (Überblick)

1.6.1 Unterschiede zu traditionellen Datenbanken

In relationalen und objektorientierten Datenbanken spezifiziert der Benutzer die Bedingungen, die das gesuchte Anfrageobjekt erfüllen muß. In SQL werden deklarative Anfragen formuliert, die z.B. bestimmte Attributswerte in den Ergebnistupeln fordern.

In Multimedia-Datenbanken sind exakte Anfragen nach bestimmten Attributswerten (z.B. in Metadaten) ebenfalls möglich, stellen aber in der Praxis eher die Ausnahme dar. Typischerweise möchten Benutzer eher auf der Basis von Ähnlichkeit recherchieren. Für die Spezifikation von Anfragen gibt es folgende Möglichkeiten:

- Ein Multimedia-Objekt wird vom Benutzer zur Verfügung gestellt (z.B. durch Eingabe einer URL oder durch Angabe einer Datei etc.). Das System sucht in der Datenbank nach Objekten, die zu dem vorgegebenen Objekt möglichst ähnlich sind (je nach Typ des Objektes sind verschiedene Ähnlichkeitsmaße denkbar)
- Das System stellt z.B. einen einfachen grafischen Editor zur Verfügung, mit dem man das Anfrageobjekt skizzieren kann.

In beiden Fällen ist keine einfache Ja/Nein-Entscheidung möglich, ob ein Datenbankobjekt zum Anfrageergebnis gehört. Oft sind die einzelnen Datenbankobjekte “mehr oder weniger gute” Ergebnisse der Anfrage.

1.6.2 Ähnlichkeitsmodelle

Erste und wichtigste Frage für die Recherche nach ähnlichen Objekten ist die nach einem für die Anwendung geeigneten Ähnlichkeitsmaß. Für die unterschiedlichen Medientypen wurden z.T. sehr viele verschiedene Ähnlichkeitsmaße entwickelt, die unterschiedliche Aspekte der Medienobjekte in den Vordergrund stellen.

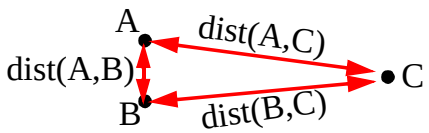
Bei Farbbildern basieren z.B. sehr viele Ähnlichkeitsmodelle auf Farbhistogrammen, d.h. es werden diejenigen Bilder als ähnlich definiert, die die selben Farben mit ähnlichen relativen Häufigkeiten enthalten. Andere Ähnlichkeitsmodelle für Farbbilder versuchen eher, die Bilder Pixel für Pixel zu vergleichen.

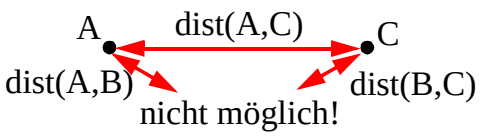
Den ersten Schwerpunkt unserer Vorlesung werden deshalb Ähnlichkeitsmodelle für verschiedene Medien bilden, wie z.B.

- Texte und andere Reihen kategorischer Werte wie z.B. Gensequenzen
- Zeitreihen wie z.B. elektrische Signale, Aktienkurse, Fieberkurven
- Bilder
- Geometrische und grafische Objekte wie z.B. CAD-Zeichnungen
- Dreidimensionale Objekte wie z.B. Bauteile, Moleküle
- komplexe Graphen wie z.B. elektronische Schaltungen

Ein häufig gewählter Ansatz versteht die Ähnlichkeit als eine *Distanz* zwischen zwei Objekten. Oft bildet die Distanzfunktion eine *Metrik*,

- d.h. die Distanzfunktion ist 0, wenn die Objekte *gleich* sind.
- Eine metrische Distanzfunktion bei *verschiedenen* Objekten immer *positiv.*
- Symmetrie: dist(*A,B*) = dist(*B,A*)
- Es gilt die *Dreiecksungleichung*, d.h. zu zwei sehr unähnlichen Objekten kann es kein drittes Objekt geben, das zu den beiden anderen wiederum sehr ähnlich ist.

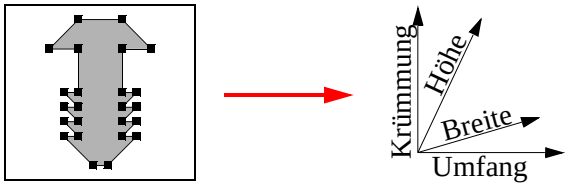




1.6.3 Feature-basierte Ähnlichkeit

Ein sehr wichtiger Spezialfall metrischer Distanzfunktionen ist die feature-basierte Ähnlichkeit, die die Multimedia-Objekte nicht nur in einen metrischen Raum sondern in einen Vektorraum überführt. Der featurebasierte Ansatz extrahiert aus den Multimedia-Objekten mehrere charakterisierende Eigenschaften als numerische Werte und erzeugt somit

Vektoren eines mehrdimensionalen (oft: hochdimensionalen) Vektorraums (genannt Feature-Raum).



Die Ähnlichkeit zweier Objekte ist dann definierbar als Abstand der Vektoren, wobei man auch hier verschiedene Abstandsmaße definieren kann, z.B.

- euklidische Distanz $\text{dist}(p, q) = \sqrt{\sum_{0 \le i < d} (p_i - q_i)^2}$
- gewichtete euklidische Distanz $\text{dist}(p, q) = \sqrt{\sum_{0 \le i < d} w_i \cdot (p_i - q_i)^2}$
- quadratische Formen $\text{dist}(p, q) = \sqrt{\sum_{0 \le i < d} \sum_{0 \le j < d} (p_i - q_i) \cdot w_{i,j} \cdot (p_j - q_j)}$
- Manhattan-Distanz $\text{dist}(p, q) = \sum_{0 \le i < d} |p_i - q_i|$
- Maximum-Distanz $\text{dist}(p,q) = \max \{|p_i - q_i|\}$

1.6.4 Algorithmen zur Ähnlichkeitssuche

Featurebasierte Ähnlichkeitsmaße sind deshalb besonders wichtig, weil es für Vektorräume effiziente Techniken zur Indexierung und Anfragebearbeitung gibt. Einige dieser Techniken werden in Kapitel 3 der Vorlesung behandelt. Insbesondere sind dies Algorithmen für

- **Range Queries (Bereichsanfragen):**
  vorgegeben ist ein Anfragevektor *q* und eine maximale Distanz ε
  alle Datenbankvektoren mit einem Abstand von höchstens ε werden ermittelt
- **Nearest-Neighbor-Queries und *k*-Nearest Neighbor Queries:**
  vorgegeben ist ein Anfragevektor und die Anzahl *k* von gesuchten Datenbankvektoren
  die *k* zum Anfragevektor nächstgelegenen Datenbankvektoren werden ermittelt
- **Ranking Queries:**
  die Datenbankvektoren können in aufsteigendem Abstand vom Anfragevektor abgerufen werden

Wir werden bei diesen Algorithmen sowohl unorganisierte Dateien (d.h. einen *sequential scan*) als auch multidimensionale Indexstrukturen (*R-Baum*) zugrunde legen. Außerdem werden wir Techniken zur *mehrstufigen Anfragebearbeitung* behandeln, bei denen zunächst in einem *Filterschritt* aus dem Index *Kandidaten* effizient ermittelt werden, die potentiell Anfrageergebnisse sind. Mit Hilfe eines *Verfeinerungsschritts* wird abschließend geklärt, welche Kandidaten tatsächliche *Treffer* der Anfrage sind.

1.6.5 Hochdimensionale Räume

Es stellt sich heraus, daß R-Bäume und andere Indexstrukturen, die im Hinblick auf Geoinformationssysteme und räumliche Datenbanken (2d/3d) entwickelt wurden, für hochdimensionale Featurevektoren schlecht geeignet sind. Bereits bei moderaten Dimensionen wie z.B. 5-8 verschlechtert sich das Leistungsverhalten zusehends. Dies bezeichnet man als *curse of dimensionality.* In Kapitel 4 werden wir einerseits die Frage untersuchen, woran das liegt, andererseits einige spezialisierte Indexstrukturen behandeln, die speziell mit Hinblick auf diese Probleme entwickelt wurden. Im einzelnen behandeln wir:

- Indexstrukturen für hochdimensionale Vektorräume
  - — X-tree, TV-tree, pyramid technique, usw.
- Kostenmodelle für hochdim. Indexstrukturen und die Optimierung der Indexstrukturen
  - — Seitengrößenoptimierung
  - — Optimierung des Indexdurchlaufs