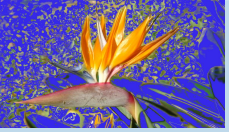
Vorlesung: Bio-Datenbanken

Kapitel 4: Ähnlichkeit nach Abstand

Dr. Dieter Sosna

10. Dezember 2007



Kapitel 4: Ähnlichkeit (Abstand)

Allgemeines

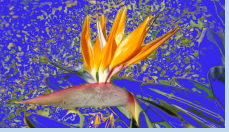
Mathematischer Abstandsbegriff

Mengen

Zeichenketten

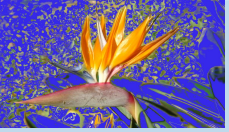
Vektoren

Abstand von Bildern



Zwischenstand

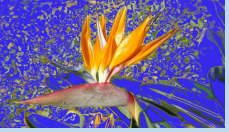
- Wichtiger Aspekt bei Datenintegration:
  Finden von Daten, die sich verschiedenen Quellen befinden und sich auf das gleiche Objekt der Welt oder des theoretischen Modells beziehen.
  Deshalb Grundfunktionen:
  Test auf Gleichheit bzw. Ähnlichkeit.
- Der mathematische Ähnlichkeitsbegriff (Äquivalenzbegriff) ist nur in wenigen Beispielen vertreten.
- Meßfehler u.ä. bedingen einen schwächeren Begriff.



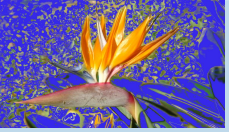
Einfache Objekte

Definition: Komplexes Objekt g.d.w. in Konstruktionsvorschrift der Klasse wird eine der folgenden Aggregationen *List*, *Array*, *Set*, *Bag*, *Tupel* benutzt.

- Zunächst Ähnlichkeit einfacher (nicht komplexer) Objekte.
  durch Abstandberechnungen
- Ähnlichkeit von Mengen, Arrays (Vektoren), Zeichenketten.
  Individuelle, semantisch bedingte Ähnlichkeitsdefinition
  deshalb mehrere Lösungen möglich. Realisierungen durch Strukturvergleiche, Inhaltsvergleiche, Mischformen.



Mathematischer Abstandsbegriff

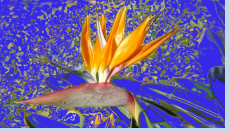


Motivation der Abschwächung

- Grund Ausgleich von Meßungenauigkeiten
- ziemlich ähnlich ($\sim^z$): Gegeben eine Zahl $\varepsilon > 0$. Zwei Dreiecke D_1, D_2 heißen ziemlich ähnlich g.d.w. sich jeder Winkel von seiner Entsprechung im anderen Dreieck höchstens ε unterscheidet:
  $\max(|\alpha_1 - \alpha_2|, \beta_1 - \beta_2|, |\gamma_1 - \gamma_2|) \leq \varepsilon$
- Beispiel: Gegeben $\varepsilon = 0,1$,
  3 Dreiecke D_1, D_2, D_3 mit jeweis passendem 3. Winkel.

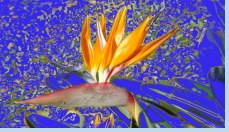
Δ	1	2	3
α	0,5	0,6	0,7
β	1	1	1

  Dann gilt $D_1 \sim^z D_2$ und $D_2 \sim^z D_3$
  aber nicht $D_1 \sim^z D_3$ Verlust der Transitivität
- *Ähnlichkeit* in zwei *homonymen* Bedeutungen:
  (geometr.) Ähnlichkeit vs. Clusterbildung
  D_1, D_2, D_3 im Cluster (um Zentrum D_2 und mit $\varepsilon < 0,1$).



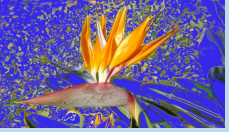
Mathematischer Abstandsbegriff

- Funktionalanalysis - Metrische Räume
  Seien $\mathcal{D}$ eine Vektorraum, ρ eine Abbildung, ρ: $\mathcal{D} \times \mathcal{D} \mapsto \mathcal{R}^+ \cup \{0\}$ mit:
  i : $\rho(x, y) \geq 0$ für $x, y \in \mathcal{D}$, $\rho = 0 \leftrightarrow x = y$.
  ii: $\rho(x, y) = \rho(y, x)$, $x, y \in \mathcal{D}$ (Symmetrie)
  iii: $\rho(x, y) \leq \rho(x, z) + \rho(z, y)$, $x, y, z \in \mathcal{D}$ (Dreiecksungleichung),
  $\rho(., .)$ heißt eine Metrik auf $\mathcal{D}$
- ohne die Bedingung $\rho = 0 \leftrightarrow x = y$: Pseudometrik
- Informatik: $\mathcal{D}$ sei (nur) eine Menge.
- Zu einer Menge kann es mehrere, verschiedene Abstandsdefinitionen geben ($\rightarrow$ verschiedene Räume)
- Sei $\mathcal{B}$ ein normierter Raum mit der Norm $\|.\|$, dann ist $\rho(x, y) = \|x - y\|$ eine Metrik.
- Nicht aus Norm erzeugt: *Diskrete Metrik*: $\rho = 0 \leftrightarrow x = y, \rho = 1$ sonst.



Beispiele normierter Räume

- $\mathcal{D}$ Menge der über einem abgeschlossenen Intervall I stetigen Funktionen f:
  $\|f\| = \max_{x \in I}(|f(x)|)$.
- L_1, L_p: $\mathcal{D}$ Menge der messbaren Funktionen über einem abgeschlossenen Intervall I mit
  $\int_I |f|^p dx < \infty, 1 \leq p < \infty$, fest, $\|f\| = \int_I |f|^p dx, 1 \leq p < \infty$.
- L_∞: $\mathcal{D}$ Menge der messbaren Funktionen über einem abgeschlossenen Intervall I mit
  ess $\sup_{x \in I}(|f(x)|) < \infty$, $\|f\| =$ ess $\sup_{x \in I}(|f(x)|)$.



Beispiele - diskreter Fall

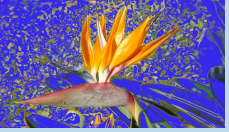
Folgen:

- l_1, l_p: $\mathcal{D}$ Menge der Folgen $\tilde{a} = \{a_i\}_{i=1}^{\infty}$ mit
  $\sum_{i=1}^{\infty}(|a_i|^p) < \infty$, $1 \leq p < \infty$, fest, $\|a\| = \sum_{i=1}^{\infty}(|a_i|^p)$.
- l_∞: $\mathcal{D}$ Menge der Folgen $\tilde{a} = \{a_i\}_{i=1}^{\infty}$ mit
  $\max_i(|a_i|) < \infty$, $\|a\| = \max_i(|a_i|)$.

Endlich viele Folgenglieder: $\tilde{a} = \{a_i\}_{i=1}^{m}$

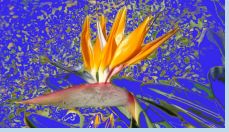
- $\mathcal{D}$ Menge der Folgen $\tilde{a} = \{a_i\}_{i=1}^{m}$ mit
  $\|a\| = \sum_{i=1}^{m}(|a_i|^p)$, $1 \leq p < \infty$, fest.
  $p = 1$ führt auf die Manhattan-Metrik,
  $p = 2$ auf die Euklidische.
- $\mathcal{D}$ Menge der Folgen $\tilde{a} = \{a_i\}_{i=1}^{m}$ mit
  $\|a\| = \max_i(|a_i|)$.

Freiwillige Übungsaufgabe: Skizzieren Sie für $m = 2$ das Aussehen des Einheitskreises in Abhängigkeit von p.



Endlich viele Folgenglieder

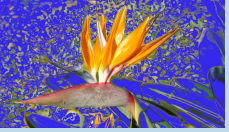
- $\tilde{a} = \{a_i\}_{i=1}^m$ kann als Vektor der Dimension m gelten.
- die Manhattan-Norm, die euklidische Norm und die Maximum-Norm sind äquivalent, d.h.
  es gibt Konstanten $c_1, c_2 \in R$, mit denen eine Norm die andere nach oben und nach unten abschätzt:
  $$c_1 \|.\|_1 \leq \|.\|_2 \leq c_2 \|.\|_1$$
  (Beweis: Ausrechnen.)
  M.a.W.: man kann häufig zu einer vorteilhafteren Norm gehen,
  (beispielsweise ist die Manhattannorm vielfach einfacher zu berechnen als die euklidische Norm).
  ÜA:
  Die Konstanten hängen von m ab!
  Berechnen Sie die Konstanten für $m = 2$ und für $m = 3$.



Distanzfunktionen mit Gewichten

Sei $\mathcal{A}$ eine positiv semidefinite m-reihige Matrix.
$x, y \in \mathcal{R}^m$.

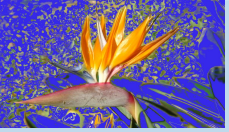
- Gewichtete Distanzfunktion:
  $$\rho(\mathcal{A}; x, y) = \left((x-y)^T \mathcal{A}(x-y)\right)^{1/2}$$
  Anwendung: Modellierung eines Farbkreisesgleicher Helligkeit (Empfindlichkeit des Auges ist farbabhängig).
- Sonderfall:
  $\mathcal{A}$ hat Diagonalgestalt: Euklidische Distanz mit Gewichtung der Achsenrichtungen.
  - Beispiel: $\mathcal{A} = (a_{i,j})_{1=1,j=1}^{m,m}$, mit $a_{i,j} = 0$ für$j \neq j, a_{i,i} = 1/i$
    (Unterschiede werden umso schwächer bewertet, je höher der Index)
  - Wählt man für $\mathcal{A}$ die Einheitsmatrix, erhält man die euklidische Distanz.
    $$\left((x-y)^T(x-y)\right)^{1/2} = \left(\sum_{i=1}^{m}(x_i - y_i)^2\right)^{1/2}$$



Triviales Beispiel für Abstandsfunktion:
Betrag: seien a, b zwei reelle Zahlen, euklidischer Abstand:

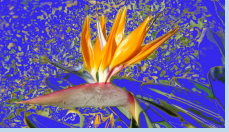
$$\rho(a,b) = \|a-b\| = ((a-b)^2)^{1/2} = |a-b|$$

(Metrik durch Norm erzeugt, beachten Sie $(a^2)^{1/2} = |a|, a \in \mathcal{R}$.
Nachweis der Eigenschaften einer Metrik : freiwillige ÜA.

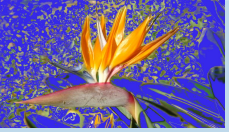


Beispiele für Abstandsmaße

- Aufzeigen von Beispielen für die Komplexbildenden Grundkonstruktionen der Informatik.
  Mengen, Vektoren (Zeichenketten)
- Varianten, Kombinationen



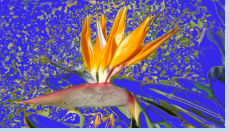
Mengen



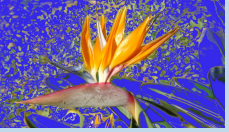
Hausdorffdistanz

- Warnung vor der scheinbaren Triviallösung.
- Abstand zweier kompakter Mengen $\mathcal{A}, \mathcal{B}$ eines metrischen Raumes $\mathcal{R}$, Metrik $d(.,.)$ kompakte M. im metr.Raum: Grenzwert jeder konverg. Folge gehört zur Menge.
  - gerichteter Abstand:
    $d1_H(\mathcal{A}, \mathcal{B}) = \max(\sup_{a \in \mathcal{A}} \inf_{b \in \mathcal{B}} d(a, b)$
  - Hausdorff-Distanz:
    $d_H(\mathcal{A}, \mathcal{B}) = \max(\sup_{a \in \mathcal{A}} \inf_{b \in \mathcal{B}} d(a, b), \sup_{b \in \mathcal{B}} \inf_{a \in \mathcal{A}} d(a, b))$
  - Verbal:
    Zwei Mengen haben eine HD von höchstens r voneinander, g.d.w. jeder Punkt einer Menge ist innerhalb eines Abstandes r von einem Punkt der anderen.

S.auch unten: Ähnlichkeit nach Inhalt.
Beachten Sie: Gleichheit von Mengen ist durch gleichen Inhalt definiert:
$\mathcal{A} = \mathcal{B}$ g.d.w. $\mathcal{A} \subseteq \mathcal{B} \wedge \mathcal{B} \subseteq \mathcal{A}$.

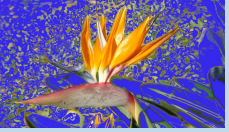


Zeichenketten



Zeichenketten

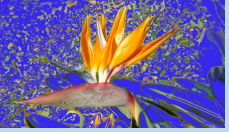
- Zeichenkette:= Liste von Elementen (Buchstaben) aus einer Grundmenge (Alphabet).
  Ggf. auch als Array ansprechbar oder als spezielle Vektoren
- Abstanddefinitionen:
  Typ 1: spezielle für Zeichenketten und
  Typ 2: allgemeine für Vektoren.
  Typ 3: aus den Zeichenketten neue Objekte ableiten, für diese neue Abstanddefinitionen geben und diese als Abstand der Zeichenketten definieren.



Hamming-Distanz

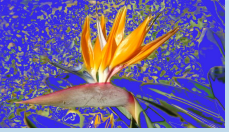
Richard W. Hamming. Error Detecting and Error Correcting Codes,
Bell System Technical Journal 26(2):147-160, 1950.

- Gegeben:
  Alphabet $\mathcal{A}$, 2 Zeichenketten $a = \{a_i\}_{i=1}^n, b = \{b_i\}_{i=1}^n$ der Länge n.
  $d_H(a,b) = \sum_{i=1, a_i \neq b_i}^{n}(1)$
  d_H ist eine Metrik auf der Menge der Zeichenketten der Länge n.
- Beispiel:
  $\mathcal{A} = \{0,1\}$, Zeichenkette: Binärzahlen der Länge n
  $d_H(a,b) =$ Anzahl der 1-Zeichen in a xor b.
  Darstellung des Übergangs von a nach b als Kantenfolge in einem n-dimensionalen Hyper-Würfel.
  Manhattan-Abstand
  Beispiele: http://en.wikipedia.org/wiki/Hamming_distance



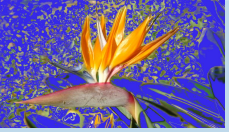
Hamming-Distanz (2)

- Mögliche Anwendung : Fehlerkorrektur
  Voraussetzung: es gibt eine Menge der korrekten Zeichenketten $\mathcal{K}$
  Falls Zeichenkette $a \notin \mathcal{K}$ suche in $\mathcal{K}$ nach Zeichenkette mit dem kleistem Abstand zu a und ersetze damit a.
- Probleme:
  Eindeutigkeit der Lösung des Minimalproblems evt. nicht gegeben,
  die gefundene Lösung muß nicht korrekt sein, insbesondere bei mehrfachen Fehlern, (Sprachwissenschaften Ergänzung durch andere Heuristiken, Häufigkeitsannahmen u.s.w.
  ...



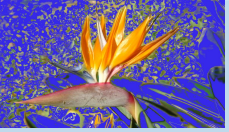
Levenstein-Distanz

- Auch: edit-distance
- **Definition:** Gegeben zwei Zeichenketten $x = \{x_i\}_{i=1}^{n}$, $y = \{y_j\}_{j=1}^{m}$.
  Grundoperationen mit Gewicht
  insert(x, c, l): fügt in Zeichenkette x das Zeichen c an der Position l ein. Gewicht g_i.
  delete((x, l): löscht in Zeichenkette x das Zeichen an der Position l. Gewicht g_d.
  replace(x_l, c, l): ersetzt in Zeichenkette x das Zeichen an der Position l durch c. Gewicht g_r.
  Gesucht: eine Folge von Grundoperationen minimalen Gesamtgewichts d (= Summe der Gewichte), die x in y überführt.
  Das **Gesamtgewicht einer Minimalfolge ist die Levenstein-Distanz** von x und y.



Levenstein-Distanz (Verallgemeinerungen)

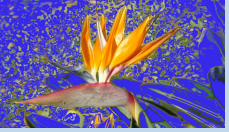
- Gültigkeit einer Dreiecksungleichung für Gewichte für Operationen an einer Position - jede Position nur einmal bearbeitet.
- Die Gewichte können abhängen vom Zeichen (sowohl dem zu ersetzenden und dem ersetzenden) (unsymmertr. Metriken, symmetrisierbar)
- Verallgemeinerung auf Baumstrukturen



Levenstein-Distanz (Berechnung)

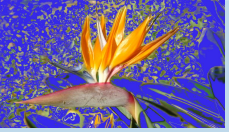
- Idee: Berechnung der Distanz aller möglichen Präfix-Paare der zwei Zeichenketten x, Y.
- $x = ua$, $y = vb$.

$$g(ua, vb) = \min \begin{cases} g_d(x, .) + g(u, vb) & - \text{ Loeschen von a} \\ g_i(., b, .) + g((ua, v) & - \text{ Einfuegen von b} \\ g_r(a, b, .) + g(u, v) & - \text{ Ersetzen a durch b} \end{cases}$$

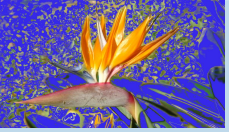


Levenstein-Distanz für Baumstrukturen

- Definition: Ein *Baum* besteht aus einem Knoten und einer daran angehängten, geordneten Folge disjunkter Bäume. Eine solche Folge heißt *Wald*.
- Grundoperationen: (jeweils mit Kosten zu versehen)
  Ersetzen eines Knotens (ändert Baumstruktur nicht)
  Einfügen eines Knotens (verschiebt den neuen Wald)
  Löschen eines Knotens (verschiebt den Wald).
- Gegeben zwei Wälder $\mathcal{F}, \mathcal{G}$. Sei $\mathcal{X}$ die Menge aller Folgen von Grundoperationen, deren Hintereinanderausführung $\mathcal{F}$ in $\mathcal{G}$ überführt. Die Editier-Distanz $d(\mathcal{F}, \mathcal{G})$ ist das kleinste Gesamtgewicht eines Elenents aus $\mathcal{X}$
- Algorithmen: Tai - 1979: $O(n^6)$, Zhang-Shasha - 1989:$O(n^4)$, Klein - 1998: $O(n^3 \log n)$.
  Forschungsgegenstand. (2004, 2005, ...)

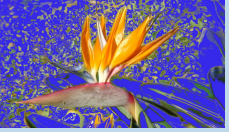


Vektoren

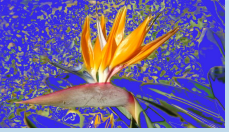


Ähnlichkeit von Vektoren

- vgl.: Math. Abstandsbegriff - normierte Räume sind Vektorräume.
- Hilberträume $\mathcal{R}$: Skalarprodukt $(.,.)$ (verträglich mit Norm)
- $d(x,y) = 1 - \frac{|(x,y)|}{\|x\| \times \|y\|}$ für $x, y \in \mathcal{R}$.
  Anschaulich im $\mathcal{R}^2$: $d(x,y) = 1 - |\cos(x,y)|$
  d.h. Abstand gering - fast gleiche Richtung.
  Verallgemeinerung: ohne Betrag
- Ähnlichkeit nach Tanimoto
  $d(x,y) = 1 - \frac{(x,y)}{\|x\|^2 + \|y\|^2 - (x,y)}$ für $x, y \in \mathcal{R}$.
  Vergleiche von Molekülstrukturen in Bio-DB und Chemie-DB: Fingerprint - Bitkette
  Führt zu anderem Ähnlichkeitsbegriff - Übergang zu inhaltsbezogener Ähnlichkeit.

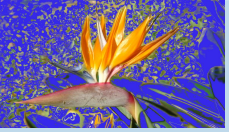


Abstand von Bildern



Ähnlichkeit von Bildern

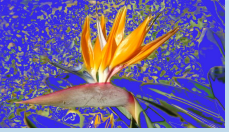
- Formale Daten: Größe, Kodierung, Exif-Daten (Photo)
  Farbwerte an ausgewählten Koordinaten (Gen-array).
- Inhaltsbezogene Verschlagwortung (teuer)
- Ermittlung typischer Werte hinsichtlich Farben (Sonnenuntergang, ...),
  Farbverteilungen, ...
- Niedere Koeffizieten der Fourier-Transformierten
  (JPEG, MP§ bei Ton)



Ähnlichkeit von Bildern (2)

Charles E. Jacobs: Fast Multiresolution Image Querying. Proc. SIGGRAPH 1995.

- aus Inhalt charakteristische Daten errechnet:
  Farbmodell YIQ, Wavelettransformation (Haar Wavelets)
- Idee d.Metrik: gewichtete L_1-Norm von bearbeiteten WL-Koeffizienten der Bilder Q, T für jeden Kanal des Farbmodells:
  $\|Q,T\| = w_{0,0}|Q(0,0) - T(0,0)| + \sum_{i,j} w_{i,j}|\tilde{Q}(i,j) - \tilde{T}(i,j)|$
- Praktische Metrik noch vereinfacht (Symmetrieverlust)- ist dann im math Sinn keine Metrik.
- u.a. Vergleiche zwischen Kinderzeichnungen und Photographien möglich.



Ähnlichkeit von Bildern (2)

Charles E. Jacobs: Fast Multiresolution Image Querying. Proc. SIGGRAPH 1995.

- aus Inhalt charakteristische Daten errechnet:
  Farbmodell YIQ, Wavelettransformation (Haar Wavelets)
- Idee d.Metrik: gewichtete L_1-Norm von bearbeiteten WL-Koeffizienten der Bilder Q, T für jeden Kanal des Farbmodells:
  $\|Q,T\| = w_{0,0}|Q(0,0) - T(0,0)| + \sum_{i,j} w_{i,j}|\tilde{Q}(i,j) - \tilde{T}(i,j)|$
- Praktische Metrik noch vereinfacht (Symmetrieverlust)- ist dann im math Sinn keine Metrik.