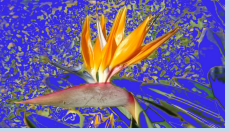
Vorlesung: Bio-Datenbanken

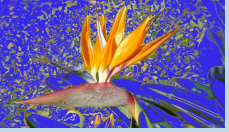
Kapitel 5: Ähnlichkeit nach Inhalt

Dr. Dieter Sosna

11. Januar 2008

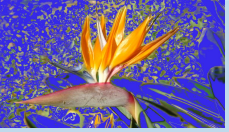


Kapitel 5: Ähnlichkeit nach Inhalt



Zwischenstand

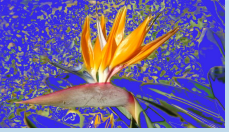
- Wichtiger Aspekt bei Datenintegration:
  Test auf Gleichheit bzw. Ähnlichkeit.
- Der mathematische Ähnlichkeitsbegriff (Äquivalenzbegriff) ist nur in wenigen Beispielen vertreten.
- Ähnlichkeit bei geringem Abstand



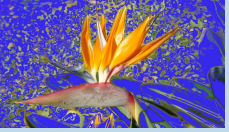
Komplexe Objekte

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

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

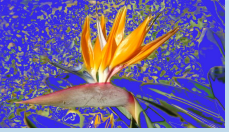


Inhaltsbasierte Ähnlichkeit



Inhaltsbasierte Ähnlichkeit

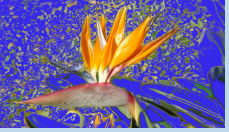
- Idee: Zwei (Objekt-) Klassen sind gleich g.d.h. sich die Instanzen der einen Klasse eineindeutig auf die Instanzen der anderen Klasse abbilden lassen.
- Theoretisch evt. abzählbar viele Instanzen - praktisch nur endl. viele, selten vollständig, d.h. eineindeutige Abbildung nur an einzelnen Zustanden verifizierbar - keine absolute Sicherheit.
- Probleme (bei Automatisierung):
  Massendatenverarbeitung
  Fehlende Metainformationen (z.B. über Struktur, Semantik)
  Beispiel: $Literatur(\{Autor\}, Titel, Verlag)$
  $Literatur(Verlag, \{\{Autor\}, Titel\})$
  $Literatur(Autor, \{Titel, Verlag\})$
  Mehrere Varianten der interen Struktur von Autor und Verlag.
  Beispiel: $Autor($ < Text > $)$
  $Autor(Name$ (m. Zusätzen$), \{Vornamen\}$
  $Autor(Name, Zusatz, \{(Vorname|Initiale)\})$
  Zusätze: von, de, ...; aber franz.: $DeFries$.



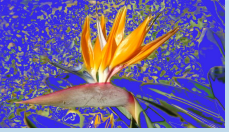
Einfache Objekte

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

- Ähnlichkeit komplexer Objekte $\Rightarrow$ Schemaintegration.
  Zunächst Ähnlichkeit einfacher (nicht komplexer) Objekte.
- Grundannahme: Zu einer Klasse gibt es eine Menge von Instanzen.
  Zustand einer Instanz = Wertebelegung.
  a) Aus der Menge der Werte einer Instanz einer Klasse werden Parameter berechnet und mit theoretischen Werten verglichen $\rightarrow$ Zugehörigkeit.
  b) Aus zwei Mengen von Werten von Instanzen werden Parameter berechnet und miteinander verglichen $\rightarrow$ Zusammengehörigkeit.
- Grundproblem: Der aktuelle Zustand einer Instanzmenge ist i.a. nur ein möglicher Zustand. Deshalb haben alle Aussagen, die nicht auf allen möglichen Zuständen basieren, das nicht zu vernachlässigende Risiko, falsch zu sein.
  Reduzierung: verschiedene Überprüfungen.

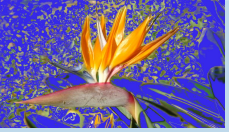


Statistische Verfahren



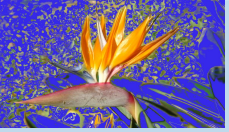
Statistische Tests - Beispiele

- Parametertests
- Parameterfreie oder nichtparametrische Tests:
  χ^2-Test: Test, ob Stichprobe einer (zuvor angenommenen) W-Verteilung F folgt.
  Kolmogorow-Smirnow-Test: Test auf Übereinstimmung zweier Wahrscheinlichkeitsverteilungen oder
  Test, ob Stichprobe einer (zuvor angenommenen) W-Verteilung folgt.
  nichtparametrischer Test, sehr stabil, für stetige, diskrete, rangskalierte Merkmale
  sehr flexibel nutzbar $\Rightarrow$ evt. nicht sehr scharf.



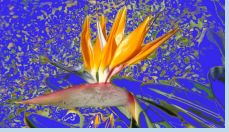
Regressionsanalyse

- Zwischen Attributen einer Instanz bestehe funktionaler Zusammenhang. Für die Daten jeder Menge von Instanzen werden Parameter bestimmt. Stimmen die Parameter für verschiedene Datenmengen überein, ist das ein Hinweis auf gleiche bzw. gleichartige Objekte.
- Aus endlich vielen Werten kann die Funktion nicht absolut sicher bestimmt werden.
  $\Rightarrow$ bei ungünstiger Datenlage falsch positive Ergebnisse.
  Beispiel: Gegeben eine Menge von Paaren (x,y), $45 < x < 55$ von Meßwerten, für die theoretisch gilt y=1/x. Durch Regressionsanalyse lasst sich eine Anpassung an y=-ax+b, a,b reell oder auch an y= a/log(x)+b, a,b reell bestimmen (jeweils mit geringen Fehlerquadrat).



Clusteranalyse

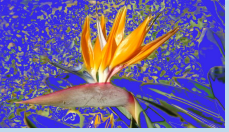
- Datenmenge wird Clusteranalyse unterworfen. Für je des Cluster werden charakteristische Werte (z.B. Clusterzentren) bestimmt. Zwei Datenmengen sind ähnlich, wenn sie die gleichen (ähnliche) Cluster bilden, d.h. wenn sie ähnliche charakteristische Werte - z.B. die Vektoren der Clusterzentren- der haben.
- weitere Abstandsmaße: s.u.



Clusterverfahren

Gruppierung von Objekten
Abbruchkriterien: z.B. Zahl der Cluster, Mindestabst. der Cluster, ...

- anhäufend:
  1. Anfangs jedes Ojekt eineigenes Cluster.
  2. Schrittweise Zusammenfassung ähnlicher Objekte bzw. Cluster zu einem neuen Cluster.
  3. Abbruchkriterum erfüllt $\rightarrow$ fertig, sonst weiter bei (2).
- teilend:
  1. Anfangs alle Ojekte in einem Cluster.
  2. Teilung der Cluster /eines Cl., so dass Abstand der Teile möglichst groß.
  3. Abbruchkriterum erfüllt $\rightarrow$ fertig, sonst weiter bei (2).
- Wahl des Abbruchkriteriums und des Schrittes (2)$\rightarrow$ verschiedene Verfahren.

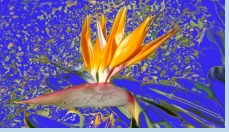


Clusterverfahren (Auswahl)

Partitionierend:

- k-means-Algorithmus (theoret. Schwächen, billig und gut)
- EM-Algorithmus
- Spektral Clusterung (Bildverarbeitung, WEB-Suche)
- Parallele Mehrfachclusterung
- ...

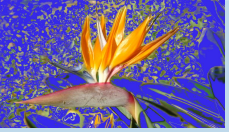
Graphentheoret. Methoden: ... Fuzzy-Clusterung



Clusterverfahren (k-mean)

k Zahl der Cluster - vorgegeben.

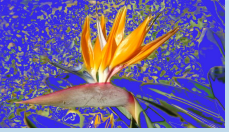
1. (Initialisierung) Auswahl von k initialen Clusterzentren
2. Jedes Objekt wird dem ihm nächsten Zentrum zugeordnet.
   Neuberechnung der Clusterzentren.
3. Ist jetzt ein Objekt falsch eingeordnet $\rightarrow$ (2).



Spektral Clusterung - Skizze

Literatur:Tutorial given at ICML 2004: Spectral Clustering.
URL: http://crd.lbl.gov/ cding/Spectral/

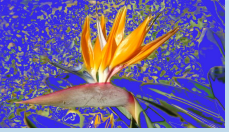
- Initialzustand: Alle Objekte in einem Cluster.
  Ähnlichkeit der Objekte i, j: $\{w_{i,j}\}$ Adjazenzmatrix der Ä.
  Schnitt S teilt Objekte in zwei Cluster A, B..
  Schnittgewicht G_S = gewichtete Kantensumme d. durchtrennten Kanten.
  Gesucht: Schnitt mit minimalem Gewicht: führt auf Eigenwertproblem für pos. semidef. Operator. Zweitkleinster Eigenwert ist die Lösung unserer Aufgabe, dazu Eigenvektor q_2.
  Mengentrennung: $A = \{i : q_2(i) < 0\}, B = \{i : q_2(i) > 0\}$,
- Da Lösung unabhängig von add. Konstante im Gewicht: Sortiere Objekte nach $q_2(i)$ und trenne in der Mitte.
- Wiederhole mit A bzw. B, wenn Abbruchkriterium nicht erfüllt oder teile weiter mit höheren EW.



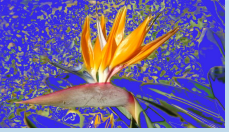
Auswertung statist. Untersuchungen

Ergebnis der stat. Untersuchungen sind neue Objekte
Ursprüngliche Objekte einander ähnlich $\Leftrightarrow$ entsprechende neue Objekte einander ähnlich.

- Regression: Koeffizienten der Regressionsgraden.
  Wann sind zwei Geraden ähnlich? ... zwei Vektoren?
- Clusteranalyse: Menge von Mengen, die um Zentroide liegen.
  Viele Ansätze:
  Ähnlichkeit von Mengen,
  Ähnlichkeit der Vektoren der Zentroide
  Abstände der Zentroide (evt. mit Gewichten)
  ...
- Probleme: ...



Mengen

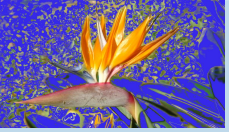


Abstandsmessungen im Inhalt

Gegeben zwei Mengen $\mathcal{A} = \{a\}, \mathcal{B} = \{b\}$ in einem metrischen Raum mit der Metrik $d(a, b)$.
Abstand der Mengen $d_M(\mathcal{A}, \mathcal{B}) =$

1. $\min_{a \in \mathcal{A}, b \in \mathcal{B}}(a, b)$ Minimaler Abstand zweier Elemente.
2. $\max_{a \in \mathcal{A}, b \in \mathcal{B}}(a, b)$ Maximaler Abstand zweier Elemente.
3. $\frac{\sum_{a \in \mathcal{A}, b \in \mathcal{B}} d(a,b)}{\text{card}(\mathcal{A})\text{card}(\mathcal{B})}$ Durchschnittlicher Abst. aller Elementpaare aus...
4. $\frac{\sum_{a,b \in \mathcal{C}, \mathcal{C} = \mathcal{A} \cup \mathcal{B}} d(a,b)}{\text{card}(\mathcal{C})}$ Durchschn. Abst. aller Paare aus Vereinigungsm.
5. $d(\bar{a}, \bar{b})$ Abst. der Mittelwerte d. Cluster (Centroid-Abst.)
6. $\frac{d(\bar{a},\bar{b})}{1/\text{card}(\mathcal{A}) + 1/\text{card}(\mathcal{B})}$ Zunahme der Varianz beim Vereinigen von $\mathcal{A}$ und $\mathcal{B}$ (Ward'sche Methode).

Achtung: evt. Verlust der Dreiecksungleichung.

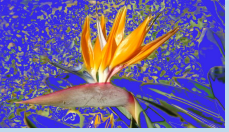


Ähnlichkeit von Mengen nach dem Inhalt (1)

Gegeben zwei Mengen $\mathcal{A} = \{a_i\}, \mathcal{B} = \{b_j\}$.
Ähnlichkeitsmaße:

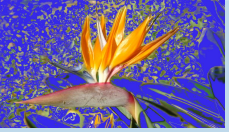
- Base: $s_{Base}(\mathcal{A}, \mathcal{B}) = \begin{cases} 1 & : \quad \mathcal{A} \cap \mathcal{B} \neq \emptyset \\ 0 & : \quad sonst \end{cases}$
- Dice: $s_{Dice}(\mathcal{A}, \mathcal{B}) = \frac{2 \times \mathsf{card}(\mathcal{A} \cap \mathcal{B})}{\mathsf{card}(\mathcal{A}) + \mathsf{card}(\mathcal{B})}$.
- Min : $s_{Min}(\mathcal{A}, \mathcal{B}) = \frac{\mathsf{card}(\mathcal{A} \cap \mathcal{B})}{\min(\mathsf{card}(\mathcal{A}), \mathsf{card}(\mathcal{B}))}$. Entsprechend: $s_{Max}(\mathcal{A}, \mathcal{B})$.

Es gilt: $s_{Max}(\mathcal{A}, \mathcal{B}) \leq s_{Dice}(\mathcal{A}, \mathcal{B}) \leq s_{Min}(\mathcal{A}, \mathcal{B}) \leq s_{Base}(\mathcal{A}, \mathcal{B})$
Definition: Zwei Mengen heißen ähnlich nach dem Maß $\mu \in \{Base, Dice, Min, ...\}$ mit dem Schwellwert s_0 g.d.w $s_\mu > s_0$.



Kommentare

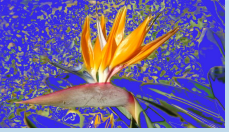
- *Dice* vernachlässigt gemeinsames Nichtenthaltensein $\rightarrow$ Faktor 2.
- Auch andere Ähnlichkeitsmaße für Bitvektoren übertragbar.
- In ähnlicher Weise auch Qualität einer Teilmengenbeziehungen über den Inhalt definierbar:
  $s(\mathcal{A} \subseteq \mathcal{B}) = \max(\frac{\mathsf{card}(\mathcal{A} \cap \mathcal{B}) - \mathsf{card}(\mathcal{A} \setminus \mathcal{B})}{\mathsf{card}(\mathcal{A})}, 0)$.
  $s = 1$ ... alle Elemente von $\mathcal{A}$ auch in $\mathcal{B}$.
  $s = 0$ ... Elemente von $\mathcal{A}$ mehrheitlich nicht in $\mathcal{B}$.



Ähnlichkeit von Mengen nach dem Inhalt (2)

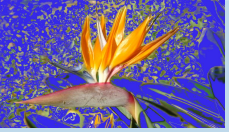
Nachweis durch Clusterverfahren
Gegeben zwei Mengen $\mathcal{A} = \{a_i\}, \mathcal{B} = \{b_j\}$

- Beide Mengen werden demselben Clusterverfahren unterworfen.
- Zwei Mengen sind hinsichtlich der Clusterbildung ähnlich, wenn es eine „wechselseitige Zuordung“der Cluster gibt, bei der
  die einander zugeordneten Cluster nach einem geeigneten Maß ähnlich sind.
  zum Beispiel:
  die gewichtete Summe der Abstände der Clustercentroide eine vorgegebene Größe nicht überschreited und
  die die relativen Häufigkeiten in den einander zugeordneten Clustern ähnlich sind.
- Ähnlichkeit der Mengen damit abgebildet auf die Ähnlichkeit des Vektors / der Menge der Cluster.

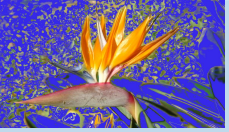


Vergleiche in Ontologien

- Minimale Eigenschaften: $Ontologie = (\{Konzept\}, \{Relation\})$
  mit „$=$“, „$\subseteq$“$\in$ { Relation }.
  Wünschenswert: auch „part-of “, „$\in$ “Elemente vom { Relation }
- Damit können Konzepte verglichen werden hinsichtlich:
  Instanzen,
  Enthaltener Spezialisierungen,
  evt. Komponenten.
- In praxi mit anderen Verfahren kombinieren: Editierdistanz, ...
  Übergang zu Schemaintegration (s. folgendes Kapitel).

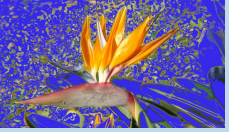


Inhaltsvergleiche anderer Strukturen



Ähnlichkeit von BIT-Vektoren

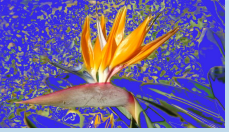
- Chemie, Biochemie: es existiert eine Liste von n Bestandteilen.
  Molekul=$(t_1, t_2, ..., t_n)$. T_j gibt an, ob das j-te Element der Liste im Molekül vorkommt (1) oder nicht (0)
- Jaccard-Koeffizient / Tanimoto-Index:
  s, t zwei Bit-Vektoren der Länge n.
  Tanimoto-Index:
  $T(s,t) = \sum_{j=1;s_j=1 \wedge t_j=1}^{n}(1) / \sum_{j=1;s_j=1 \vee t_j=1}^{n}(1)$
  Ähnlichkeit = 1 - T.
- Gemeinsames Fehlen wird ignoriert.



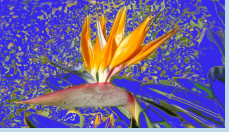
Ähnlichkeit von BIT-Vektoren II

- s, t, n wie eben,
  a Anzahl Übereinstimmungen und gleich 1,
  b Anzahl Nichtüberieinst. und $a_j = 1$,
  c Anzahl Nichtüberieinst. und $b_j = 1$,
  d Anzahl Übereinstimmungen und gleich 0.
- Auswahl:

Kovarianze	$a/n - ((a+b)/n \times (a+c)/n)$	
Jaccard	$a/(a+b+c)$	Neg. Überinst. ignor.
Dice	$2a/(2a+b+c)$	Neg. Ü. ign., pos. Ü. dopp.
Russel-Rao	$a/(a+b+c+d)$	
Sokal-Sneath	$a/(a+2(b+c)$	
Normal	$(a+d)/(a+b+c+d)$	
...		



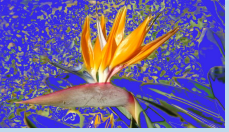
Zeichenketten



N-Gramme

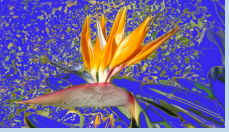
Gegeben: Zeichenketten $a, b, ...$ über einem Alphabet $\mathcal{A}$

- Ein N-Gramm x ist eine Zeichenkette über $\mathcal{A}$, die aus N Zeichen besteht. a enthält x, g.d.w. x Teilzeichenkette von a ist.
  In praxi: Bi- und Trigramme.
- Idee: Gleiche Zeichenketten = gleiche N-Gramme.
  Zwei Zeichenketten sind ähnlich, wenn sie viele gemeinsame N-Gramme haben:
  $s = 2c/(n+m)$ c Zahl der gemeinsamen Trigramme t (mit Wiederholung)
  m, n Längen der Zeichenketten
- Alternativ: Virtuell zwei Leerzeichen am Anfang und Ende: bessere Bewertung dieser Stellen.
  $s = 2c/((n+2)+(m+2))$ c Zahl der gemeinsamen Trigramme
  m, n Längen der Zeichenketten
- Metrik vom Typ 3: Mengen von N-Grammen $\Rightarrow$ Ähnlichkeit von Zeichenketten.
  Allg. Betrachtung: später.



N-Gramme (2)

- Kritisch:
  Festlegung der Ähnlichkeitsschwelle problemabhängig.
  In Berechnung gehen die Längen n, m ein.
- Anwendungen:
  Wichtiges Hilfsmittel zur Zeichenketteanalyse
  - Ähnlichkeit (s.o)
  - Anzahl der Bi- und Trigramme über einem Alphabet bekannt:
    $card(\mathcal{A})^2$,$card(\mathcal{A})^3$.
    Häufigkeitsverteilungen für Sprache oder Fachsprachen charakteristisch.
    $\Rightarrow$ Erkennung der Sprache eines Textes
    $\Rightarrow$ Zuordnung eines Textes zu Fachgebiet (zusammen mit Wortanalyse)
    $\Rightarrow$ Bestandteil der Kryptoanalyse.



N-Gramme (3)

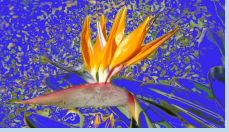
Reinhard Rapp: Die Berechnung von Assoziationen: ein korpuslinguistischer Ansatz. Hildesheim; Zürich; New York: Olms, 1996. ISBN: 3-487-10252-8 (Diss.)

- Trigrammähnlichkeit intuitiv gut.
- Einfache Buchstabendreher $\rightarrow$ unterbewertete Ähnlichkeit
- Bei gleicher (Trigramm-) Ähnlichkeit sollen mehrerer Worte zu einem Muster sollen häufigere Worte besser bewertet werden.

$$S = \frac{20c}{n+m} + \frac{2b}{n+m} + \frac{h}{10^8}$$

c : Anz. gemeins. Trigr.
n, m: Längen der Zeichenk.
h : Korpushäufigkeit des betrachteten Wortes.

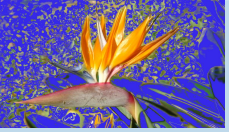
Bewertung: N-Gramme stehen an der Grenze zwischen formalen Merkmalen und inhaltsbezogenen Vergleichen



N-Gramme (4)

Andere Ähnlichkeitsmaße für Trigramme
Gegeben: Zeichenketten a, b über einem Alphabet $\mathcal{A}$
Seien $c_a(t), c_b(t)$ die Häufigkeiten des Trigramms t in den Zeichenketten a bzw. b.

- KoKS-Projekt:
  $a = \frac{\sum_t(\min(c_a(t),c_b(t)))}{\sum_t(\max(c_a(t),c_b(t)))}$
- Beziehung zum Jaccard-Maß: Ohne Beachtung der Anzahl: Auftretenen von t ist 1, Nichtauftreten 0.
- Baldwin; Tanaka (2000):
  Ersetzt man $\sum_t(\max(c_a(t), c_b(t)))$ durch das arithm. Mittel der Längen von a, b, geht a in S über (s.o.).



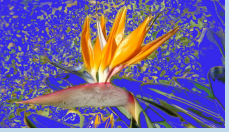
TFIDF-Maß

- n-Gramme: meist n klein, n-Gramme nicht robust gegen Wortumstellung (Dieter Sosna vs. Sosna, Dieter)
- Zerlegung der Sätze in Token (Worte) , dann z.B. Jaccard-Maß für Token
- Maß: Term-Frequenz/inverse-Dokument-Frequenz TFIDF:
  - Token t in String S erhält Gewicht $w(t, S)$:

    $$w(t, S) := \log(t \times f(t, S) + 1) \times \log(N/f(t, D) + 1)$$

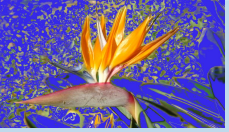
    $f(t, S)$ Häufigkeit von t in S,
    $f(t, D)$ Häufigkeit von t in allen S: $D = \cup S$.
    N Gesamtzahl aller Token.
  - Skalarprodukt der beiden Gewichtsvektoren:
    T die Vektor aller Token

    $$sim_{TFIDF}(S_1, S_2) = \sum_{t \in T}(w(t, S_1) \times w(t, S_2))/\|w(S_1\|\|w(S_2\|$$



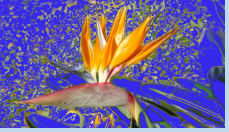
Alternativen

- Siehe auch einführendes Beispiel: Ähnliche Dreiecke
  Winkel in **Menge**: Spiegelung erhält Ähnlichkeit,
  Winkel in **Liste**: Spiegelung erhält Ähnlichkeit **nicht**.
- Es ist grundsätzlich möglich, ein- und dasselbe Objekt unter unterschiedlichen Gesichtspunkten zu sehen, entsprechend verändern sich die Eigenschaften:
  z.B Zeichenketten (auf Wortbasis)
  als Bitvektor (Alphabet, Auftreten J/N oder Anzahl)
  Zeichenkette - n-Gramme - Bitvektor oder Menge.
- Zeichenkette (Dokumentbasis): kookurente Worte (Bitvektortechnik), ...
  Übergang zu Textmining.



Sprachen

- In Bio-DB häufig Freitextfelder - (derzeit) Englisch die internationale Kommunikationssprache der Wissenschaft.
- Grundsätzliche Probleme bleiben:
  Schreibfehler,
  Synonyme, Homonyme
  Kontextabhängige Semantik
  Übersetzungsprobleme (die falschen Freunde)
- Mittel der Sprachverarbeitung zwar komplexer, häufig bessere Resultate als simple Ähnlichkeitsmaße (= komplexe Ähnlichkeitsmaße), setzt jedoch sprachliche Konstrukte (Sätze, Artikel) voraus.
- mehrsprachige Ontologien, Taxonomien wünschenswert.



Sprachprobleme

Zuordnung von Zeichenketten über Sprachgrenzen (natürl. Sprachen, Fachsprachen) - M:N.
Metawissen (Gebiet) - Ontologie d. Begriffe (Gleichheit = gl. Position in d. O.)

- (η) αναχώρηση: Abfahrt(Bahn, Schiff), Abflug (Flugzeug); 1:N
- Kontext wechseld: M:N

άγριος	:	(9 Übers.) blind (Haß, Wut), rauh (Berge, Wetter),streng (Blick)
στράβος	:	blind (nicht sehend), Syn. τυφλός
στράβος	:	blind (völlig ungebildet)
λαθραίος	:	blind (Passagier)

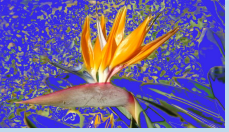
- Kontext *Verwandschaft*

(ο) γαμπρός	:	Bräutigam
(ο) γαμπρός	:	Schwager (Mann der Schwester)
(ο) κουνιάδος	:	Schwager (Bruder d. Ehefrau / d. Ehemanns)

  Kontext erweist sich als zu grob!
- Kontext in Fachsprachen:

Worte		Kontext
Quark	:	Speise; Elementarteilchentheorie
charmant, Quark	:	Elementarteilchentheorie
Farbe, Quark	:	Speise; Elementarteilchentheorie
String	:	Informatik, Kosmologie

  Kontextbestimmung durch Kookurenzen.



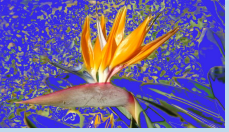
Sprachprobleme (2)

- Hohe formale Ähnlichkeit: Wein, Bein
- Kurze Präfixe oder Suffixe könen Semantik ändern.

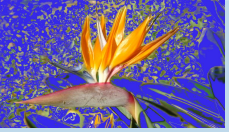
	periodisch	aperiodisch
Formale Ähnlichkeitsmaße sind hoch:	organisch	anorganisch

- Negierende Vorsilbe a-, an- (a.d.Gr.), Bsp.: Atom - ἄτομος,
  Aber: Vorsilbe an- von ανω-, πανω (ano-, pano-) = oben
  Gegenteil: κατω (kato) = unten: Anion vs. Kation, Anode vs. Kat(h)ode.
- Vorsilbe syn- (Synthese) mutiert zu sym-, syl-, sy- :
  Sympathie (Symbiose, symmetrisch), Syl-labus [Zusammenfassung], Sy-zygie [Konjunktion u. Opposition von Mond u. Sonne], Sy-stem, Sy-stole [Zusammenziehen]
- Die falschen Freunde des Übersetzers:
  Höhe formale Ähnlichkeit - unterschiedliche Semantik.

παθόλογος	(pathologos)	?	Pathologe
↑ Allgemeinmediziner		≠	↑ Spez. f.Gewebeveränderung



Konzept, beliebige komplexe Objekte auf Ähnlichkeit zu untersuchen.
Speziell überwachtes Lernen:
Bewertungsfunktion: $x \in \mathcal{X}, y \in \mathcal{Y} : (x, y) \mapsto \{w|f\}$, w g.d.w. x ähnlich y.
Lernphase: Eingabe von Paaren mit bekanntem Wahrheitswert.
Noch nicht in der Praxis. Diplomarbeit ?



Zusammenfassung

- Inhaltsbasierte Ähnlichkeit ergänzt abstandsbasierte
- Ziel: aus formalen Merkmalen semantische Ähnlichkeit erkennen.
- Semantik bestimmt meist nicht die formale Darstellung eindeutig
  und Gleiches gilt umgekehrt
  $\Leftrightarrow$ Verfahren nicht eindeutig, zu jedem Verfahren gibt es schlechte Beispiele.
- Zu komplexen Grundtypen *Menge, Vektor / Liste* sowie für Zeichenketten
  verschiedene Möglichkeiten des inhaltlichen Vergleichs.
  $\Rightarrow$ Auswahl des Verfahrens.
- Vergleich von Konzepten auf Grund eines konkreten Zustands stets unsicher,
  d.h. auch bei hoher formaler Ähnlichkeit der Zustände können die Konzepte
  unterschiedlich sein.
  $\Rightarrow$ Anwendung mehrerer Verfahren und Kombination der Ergebnisse.