Einführung in die Ähnlichkeitsmessung

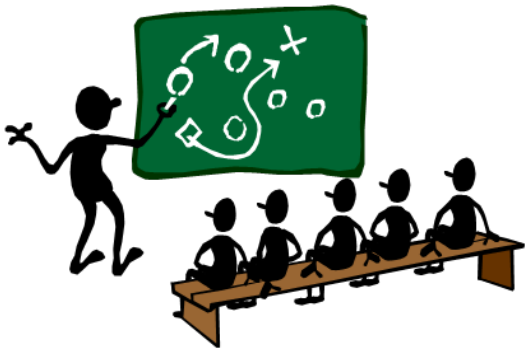
Reading Club SS 2008 – Similarity

Stefanie Sieber
stefanie.sieber@uni-bamberg.de

Lehrstuhl für Medieninformatik
Otto-Friedrich-Universität Bamberg

Agenda

- Worum geht es?
- Wann ist Ähnlichkeitsmessung möglich?
- Wie wird Ähnlichkeit gemessen?



Worum geht es?

- Ähnlichkeit (nach http://de.wikipedia.org)
  - Allgemeiner Sprachgebrauch
    - Zwei Gegenstände oder Sachverhalte weisen bei vergleichender Betrachtung gemeinsame Eigenschaften auf.
  - Mathematik
    - Geometrie

      Ähnlichkeit von Figuren bei Übereinstimmung von Winkeln und Streckenverhältnissen (Ähnlichkeitssätze zu Dreiecken – WW, SSS, SWS)
    - Lineare Algebra

      Ähnlichkeit auf Matrizen bei Verwendung unterschiedlicher Basen für dieselbe lineare Abbildung
  - Philosophie
    - Ähnlichkeit als Natureigenschaft

      Ähnlichkeit begründet in der Wiedererkennung
    - Ähnlichkeit in der menschlichen Kultur

      Dinge können erst durch Erkennung von Ähnlichkeit gruppiert und klassifiziert werden.

Vortrag Sebastian Matyas am 02. Juli 08

Worum geht es?

- Messen
  - Homomorphe Abbildung eines empirischen Relativs in ein numerisches Relativ bzw. Repräsentation eines empirischen Relativs durch ein numerisches Relativ
    - Homomorph

      strukturerhaltende Abbildung zwischen zwei mathematischen Strukturen
    - Empirisches Relativ

      Menge von Elementen und ihren Relationen zueinander, die die Art der Beziehungen der Elemente untereinander charakterisieren
    - Numerisches Relativ

      in Zahlen überführtes empirisches Relativ
  - Existenz der homomorphen Abbildung

    → Kriterium der Messbarkeit von Eigenschaften

→ Im Folgenden:

Anwendung statistischer Methoden zur eigentlichen Berechnung der Ähnlichkeit zwischen zwei Objekten

Worum geht es?

- Ähnlichkeit & Ähnlichkeitsmessung
  - **Informatik – Retrieval**
    - Ähnlichkeit im Kontext der Textsuche
    - Ähnlichkeit im Kontext der Bildsuche
    - ...
  - **Statistik - Multivariate Verfahren**
    - Allgemein: Distanzmessungen
    - Strukturprüfende Verfahren
      - Regressionsanalyse
      - Kontingenzanalysen
      - ...
    - Strukturentdeckende Verfahren
      - Faktorenanalyse
      - Clusteranalysen
      - Multidimensionale Skalierung
      - ...
    - ...

Vortrag Daniel Blank am 11. Juni 08

Vortrag Adrian Hub am 11. Juni 08

Vortrag Martin Hofmann am 02. Juli 08

Vortrag Sebastian Matyas direkt im Anschluss

Wann ist Ähnlichkeitsmessung möglich?

- Voraussetzungen für die Messung von Ähnlichkeit
  - Schwache Ordnung der Objekte
    - $(o_i, o_j) \succsim (o_k, o_l)$
  - Eigenschaften der empirischen Relation
    - X ist konnex, d.h. für alle $o_i, o_j, o_k, o_l \in A$ gilt
      - entweder $(o_i, o_j) \succsim (o_k, o_l)$
      - oder $(o_k, o_l) \succsim (o_i, o_j)$
      - oder beides.
    - X ist reflexiv, d.h. für alle $o_i, o_j \in A$ gilt $(o_i, o_j) \succsim (o_i, o_j)$
    - X ist transitiv, d.h. $\forall o_i, o_j, o_k, o_l, o_r, o_s \in A$ gilt
      - wenn $(o_i, o_j) \succsim (o_k, o_l)$ und $(o_k, o_l) \succsim (o_r, o_s)$
      - dann $(o_i, o_j) \succsim (o_r, o_s)$.
    - $\exists \mathbf{B} \subseteq \mathbf{A} \times \mathbf{A}$ mit der Eigenschaft, dass $\forall (o_i, o_j), (o_r, o_s) \in \mathbf{A} \times \mathbf{A}$ mit $(o_i, o_j) \succsim (o_r, o_s)$ ein $(o_k, o_l) \in B$ existiert, derart, dass gilt
      - $(o_i, o_j) \succsim (o_k, o_l) \succsim (o_r, o_s)$
    - ➔ Schwache Ordnungsstruktur

Wann ist Ähnlichkeitsmessung möglich?

- Voraussetzungen für die Messung von Ähnlichkeit
  - Repräsentativitätstheorem
    - Voraussetzung: Existenz der schwachen Ordnung
    - Aussage
      - Existenz der homomorphen Abbildung eines empirischen Relativs in ein numerisches Relativ
      - Beschreibung der Eigenschaften der Repräsentation
  - Eindeutigkeitstheorem
    - Voraussetzung: Existenz der homomorphen Abbildung
    - Spezifikation der Klasse der zulässigen Transformationen einer existierenden homomorphen Abbildung
    - Bestimmung des Skalentyps dieser Skala

Wann ist Ähnlichkeitsmessung möglich?

- Distanzfunktion
  - Es sei A eine nichtleere Menge. Eine Funktion d, die je zwei Objekten $\mathbf{o}_i, \mathbf{o}_j \in \mathbf{A}$ eine reelle Zahl $d(o_i, o_j)$ oder kurz d_{ij} zuordnet, heißt Distanzfunktion genau dann, wenn für alle $\mathbf{o}_i, \mathbf{o}_j, \mathbf{o}_k \in \mathbf{A}$ gilt:
    - Positivität: $d_{ij} \geq 0$
    - Identität: $d_{ij} = 0$ genau dann, wenn $o_i = o_j$
    - Symmetrie: $d_{ij} = d_{ji}$
    - Dreiecksungleichung: $d_{ij} \leq d_{ik} + d_{jk}$ **Voraussetzung für Metrik**
    - Ultrametrische Ungleichung: $d_{ij} \leq \max(d_{ik}, d_{jk})$ **Voraussetzung für Ultrametrik** (hierarchische Clusterverfahren)
  - Gruppen von Distanzfunktionen
    - Distanzfunktionen
    - Pseudo-Distanzfunktionen
    - Semi-Distanzfunktionen
    - Semi-Pseudo-Distanzfunktionen

Klasse	Si	Pos	Sym	Dreieck
Distanzfunktion	✓	✓	✓	✓
Pseudo-Distanzfunktion	✓	–	✓	✓
Semi-Distanzfunktion	✓	✓	✓	–
Semi-Pseudo-Distanzfunktion	✓	–	✓	–

Wann ist Ähnlichkeitsmessung möglich?

- Ähnlichkeitsfunktion
  - Es sei A eine nichtleere Menge. Eine Funktion s, die je zwei Objekten $\mathbf{o_i, o_j \in A}$ eine reelle Zahl $s(o_i, o_j)$ oder kurz s_{ij} zuordnet, heißt Distanzfunktion genau dann, wenn für alle $\mathbf{o_i, o_j, o_k \in A}$ gilt:
    - $s_{ij} \leq 1$
    - Identität: $s_{ij} = 1$ genau dann, wenn $o_i = o_j$
    - Symmetrie: $s_{ij} = s_{ji}$
    - Dreiecksungleichung: $s_{ij} \geq s_{ik} \cdot s_{jk}$ **Voraussetzung für Metrik**
    - Ultrametrische Ungleichung: $s_{ij} \geq \min(s_{ik}, s_{jk})$ **Voraussetzung für Ultrametrik**

Wie wird Ähnlichkeit gemessen?

- Direkte Methoden der Ähnlichkeitsmessung
  - Versuchspersonen werden angewiesen, auf Reize Reaktionen zu zeigen, die eine Bewertung oder einen Vergleich von Ähnlichkeiten beinhalten.
  - Eigene Definition von Ähnlichkeit möglich
  - Zuordnungsmethoden
    - Reaktionsinventar
    - Bewertung eines Reizpaars
  - Vergleichsmethoden
    - Direkter Vergleich von mehreren Reizpaaren
    - Maximale Informationsausschöpfung: vollständiger Paarvergleich
    - Ökonomischeres Verfahren: Triadenmethode
  - Sortiermethoden
    - Manuelles Clusterverfahren
    - Aufteilung der Reizobjekte in ähnliche Gruppen

Wie wird Ähnlichkeit gemessen?

- Indirekte Methoden der Ähnlichkeitsmessung
  - Ableitung von Ähnlichkeitsbeziehungen aus beobachtbarem Verhalten
  - Untersuchte Aspekte
    - Reizunterscheidbarkeit
    - Reizgeneralisierung
    - Reizidentifikation
  - These
    - Identifikationsfunktion zwischen Satz von Reizobjekten und Satz von Reaktionen (eindeutige Zuordnung)
  - Beispiele
    - Einschätzung von Reizobjekten auf mehrere inhaltlich festgelegte Skalen
    - Messung von Objekten mit objektiven Variablen, Festlegung der Ähnlichkeit anhand der Übereinstimmung dieser Variablen

Wie wird Ähnlichkeit gemessen?

- Distanz- und Ähnlichkeitsfunktionen bei metrischen Variablen
  - Euklidische Distanz

$$d_{ij} = \left(\sum_{l=1}^{m} (x_{il} - x_{jl})^2\right)^{\frac{1}{2}}$$

    - x_{il} = Messwert des i-ten Objekts auf der l-ten Variable
    - Distanz = Länge der direkten Verbindung zwischen den Objekten

    - Vorteile
      - Entsprechung zur räumlichen Anschauung
      - Einfache Berechnung
      - Translationsinvariant

Wie wird Ähnlichkeit gemessen?

- Distanz- und Ähnlichkeitsfunktionen bei metrischen Variablen
  - Euklidische Distanz
    - Nachteil: Interne Gewichtung
      - Keine Skaleninvarianz
      - Korrelationen zwischen Variablen
    - Gegenmaßnahmen
      - Inkaufnahme der internen Gewichtung
      - Eliminierung der Merkmalskorrelationen mittels Hauptachsenmethode
        - Betrachtung der gesamten Matrix, nicht einzelner Cluster
        - Varianz aller Faktoren = 1 (wichtige und unwichtige Faktoren gleichermaßen)
        - ➔ Betrachtung eines reduzierten Faktorraums

Wie wird Ähnlichkeit gemessen?

- Distanz- und Ähnlichkeitsfunktionen bei metrischen Variablen
  - Varianten der euklidischen Distanz
    - Quadrat der euklidischen Matrix, Durchschnittliche euklidische Distanz
    - Profilähnlichkeitsmaß
      - Transformation der euklidischen Distanz in Ähnlichkeitsfunktion
    - Dominanz- oder Supremumsmatrix
      - Distanz: maximale absolute Merkmalsdifferenz
    - Minkowski-Metrik

$$d_{ij} = \left(\sum_{l=1}^{m} \left(x_{il} - x_{jl}\right)^r\right)^{\frac{1}{r}}$$

      - Verallgemeinerung der euklidischen Distanz
      - Parameter r
        - r = 2: euklidische Distanz
        - r = 1: Manhattan-Metrik

Wie wird Ähnlichkeit gemessen?

- Distanz- und Ähnlichkeitsfunktionen bei binären Variablen
  - Jaccard-Koeffizient

$$s_{ij} = \frac{a}{a+b+c}$$

o_j \ o_i	1	0
1	a	b
0	c	d

    - Metrisches Ähnlichkeitsmaß
    - $0 \leq s_{ij} \leq 1$
    - Problem
      - Keine Berücksichtigung negativer Übereinstimmungen
  - Simple-Matching-Koeffizent

$$s_{ij} = \frac{a+d}{a+b+c+d}$$

    - Metrisches Ähnlichkeitsmaß
    - $0 \leq s_{ij} \leq 1$

Wie wird Ähnlichkeit gemessen?

- Distanz- und Ähnlichkeitsfunktionen bei binären Variablen
  - Entropiemaß
    - Totale Entropie

$$H_{T(i,j)} = 2(b+c)ld2$$

o_j \ o_i	1	0
1	a	b
0	c	d

      - Logarithmus Dualis
      - Deskriptives Streuungsmaß
        - Minimale Wert → gleiche Ausprägung aller Variablen
  - Euklidische Metrik
    - Entropiemaß = Vielfaches der euklidischen Distanz

$$d_{ij}^{\ 2} = \sum_{l=1}^{m}(x_{il} - x_{jl})^2 = b + c$$

$$H_{T(i,j)} = 2d_{ij}^2 ld2$$

    - → Verwendung der euklidischen Distanz für binäre Variablen

Wie wird Ähnlichkeit gemessen?

- Distanz- und Ähnlichkeitsfunktionen bei ordinalen und gemischten Variablen
  - Ordinal-skalierte Variablen
    - Niveau-Progression
      - Behandlung wie metrisch-skalierte Variablen
    - Niveau-Regression/Dichotomisierung
      - Überführung von ordinalen Variablen in binäre Variablen
        - Ordinale Variable mit t-Ausprägungen → t-1 binäre Variablen
        - Korrektur der resultierenden Variablen notwendig (Korrekturfaktor: 1/(t-1))
      - Abbildung auf nominale Variablen
        - Alle Werte unterhalb des Medians → 1
    - Problem
      - Unterstellung gleicher Intervalle bei Umrechnung

Wie wird Ähnlichkeit gemessen?

- Distanz- und Ähnlichkeitsfunktionen bei ordinalen und gemischten Variablen
  - Gemischte Variablen
    - Umrechnung
      - Niveau-Regression für höher skalierte Variablen
      - Niveau-Progression → Verfälschung der Ausgangsinformationen möglich
    - Allgemeine Ähnlichkeitsmatrix

$$s_{ij} = \frac{\sum_{l=1}^{m} s_{ijl}}{\sum_{l=1}^{m} w_{ijl}}$$

      - Zähler: Ähnlichkeitswert
      - Nenner: Wert zur Vergleichbarkeit der Variablen
        - 1 = Vergleich möglich
        - 0 = kein Vergleich möglich, keine Berücksichtigung negativer Übereinstimmung

Zusammenfassung

- Ähnlichkeitsmessung ...
  - ... ist in vielen unterschiedlichen Bereichen zu finden.
  - ... setzt eine schwache Ordnung der zu vergleichenden Objekte voraus.
  - ... kann direkt oder indirekt vorgenommen werden.
  - ... benötigt Distanz- oder Ähnlichkeitsmaße.
- Distanz- und Ähnlichkeitsmaße...
  - ... sind abhängig von der Art der zugrundeliegenden Variablen.
  - ... und ihre korrekte Auswahl sind entscheidend für die Evaluierung.
- Anwendungsfälle...
  - ... folgen in den nächsten Wochen & Vorträgen!

Vielen Dank für die Aufmerksamkeit!

Fragen..?
Diskussionsbedarf...?