Einführung in die Computerlinguistik

Textklassifikation

Referent: Peter Bücker
Dozentin: Wiebke Petersen

Wintersemester 2004/2005
Heinrich-Heine-Universität Düsseldorf

Textklassifikation: Allgemeines

- Was ist das?
  - Textklassifikation beschäftigt sich mit der Zuordnung von Texten in vordefinierte Klassen
  - Die Klassenprofile werden entweder automatisch gewonnen oder manuell erstellt (lernende oder nicht-lernende Systeme)

Textklassifikation: Allgemeines

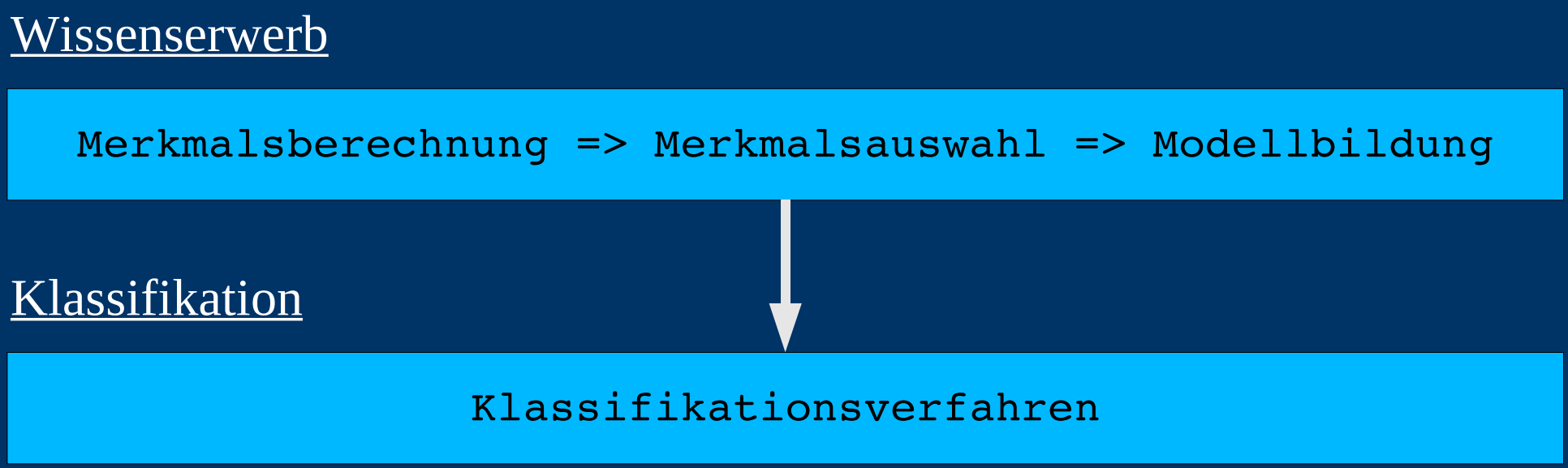
- Wozu benötigt man Textklassifikation?
  - Im Rahmen der steigenden digitalen Informationsflut ist es erforderlich geworden Algorithmen zu entwickeln, die Informationen sortieren, filtern und klassifizieren
  - Ohne diese Verarbeitung der Daten wird die Suche in Informationsquellen mit steigender Informationsmenge immer schwieriger

Textklassifikation: Allgemeines

- Welche Probleme treten dabei auf?
  - Die zu verarbeitenden Informationsmengen sind sehr groß
  - Es ist eine Unmenge an verschiedenen Klassen vorstellbar
  - Klassen können hierarchisch angeordnet sein
  - Texte können mehreren Klassen angehören
  - Natürliche Sprache ist komplex

Textklassifikation: Allgemeines

- Wie funktioniert Textklassifikation im Groben?



Textklassifikation: Allgemeines

- Wie funktioniert Textklassifikation im Groben?
  - Wissenserwerb
    - nicht-lernende Systeme
      - Experten erstellen Regeln für Trainingstexte
    - lernende Systeme
      - Merkmals-Auswahl
      - Merkmals-Gewichtung (z.B. mittels TF / IDF)

Textklassifikation: Verfahren

- Regelbasierte Verfahren
  - Klassen werden durch boolsche Terme spezifiziert
    Beispiel: `Autobahn =: Straße && (Auto && ^Fahrrad)`
  - Ein Dokument gehört genau dann zu einer Klasse, wenn die im Term spezifizierten Merkmale im Text auftauchen

Textklassifikation: Verfahren

- Regelbasierte Verfahren
  - Regeln tauchen in Form von Entscheidungsbäumen auf
  - Entscheidungsbäume können automatisch genertiert werden

Textklassifikation: Verfahren

- Regelbasierte Verfahren
  - **Vorteile**:
    - Hohe Klassifikationsqualität
    - Hohe Effizienz in der Anwendungsphase
    - Kann durch manuelle Regelanpassung optimiert werden
  - Nachteile:
    - Aufwändige Trainingsphase
    - Entscheidungsbäume werden zu spezifisch (Overfitting)
    - Keine gestufte Klassifikation

Textklassifikation: Verfahren

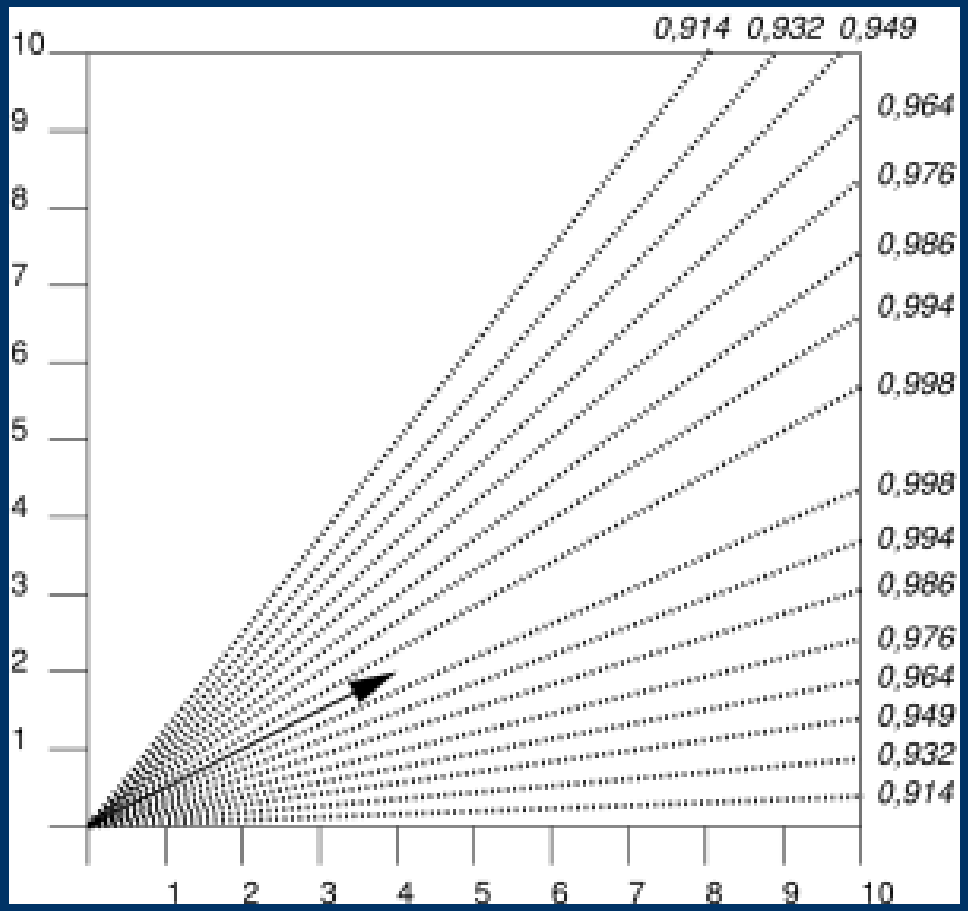
- Vektorraum-Methode
  - Eine Klasse wird durch Vektoren repräsentiert
  - Zuordnung zu einer Klasse mittels Ähnlichkeit zu einem Referenz-Vektor
  - Die Ähnlichkeit wird über ein Ähnlichkeitsmaß ermittelt

Textklassifikation: Verfahren

- Vektorraum-Methode
  - Ähnlichkeitsmaße
    - Cosinus-Maß
    - Pseudo-Cosinus-Maß
    - Skalarprodukt
    - Dice-Maß
    - Overlap-Maß
    - Jaccard-Maß
    - ...

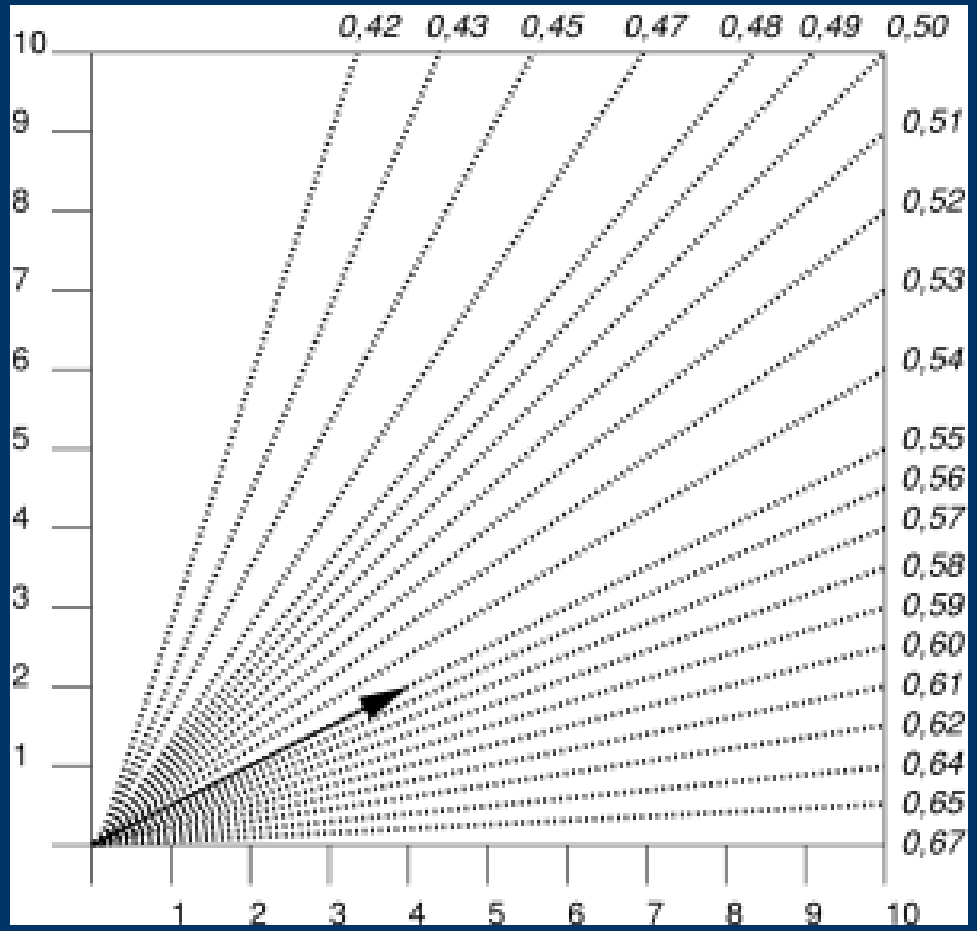
Textklassifikation: Verfahren

- Vektorraum-Methode
  - Cosinus-Maß



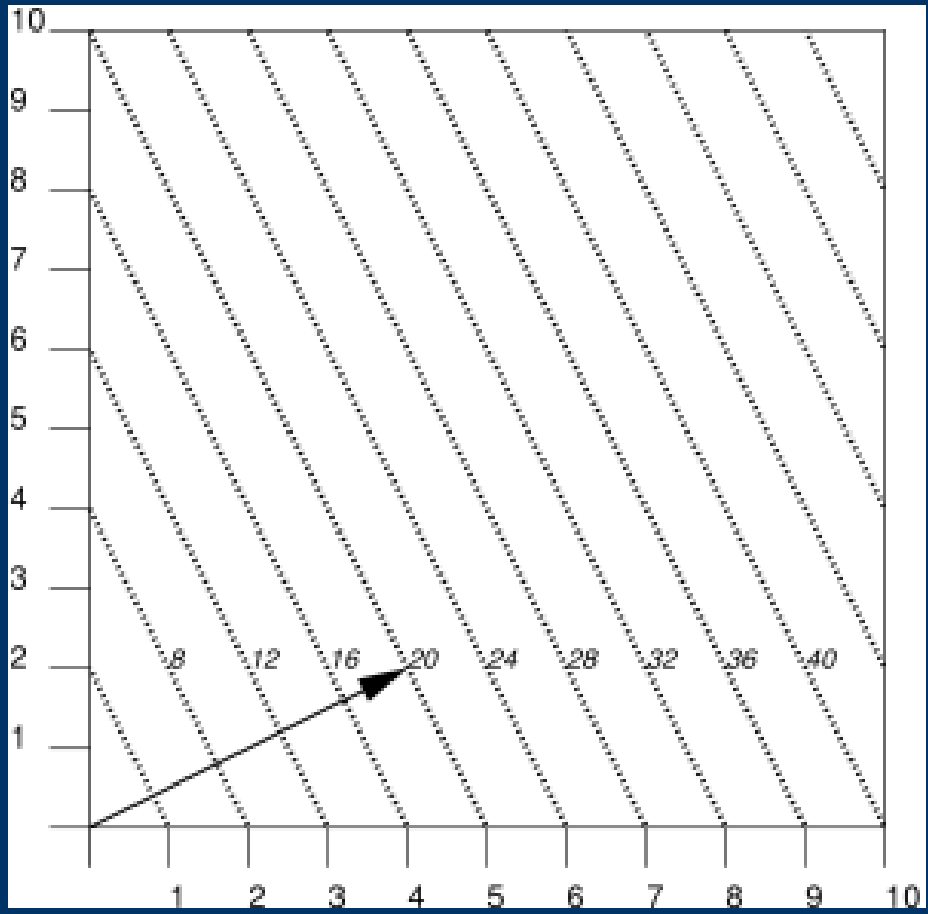
Textklassifikation: Verfahren

- Vektorraum-Methode
  - Pseudo-Cosinus-Maß



Textklassifikation: Verfahren

- Vektorraum-Methode
  - Skalarprodukt



Textklassifikation: Verfahren

- Vektorraum-Methode
  - Vorteile:
    - Gestufte Ergebnisse, keine Ja-/Nein-Antwort
    - Schnelle Generierung der Vektoren
    - Höhere Abstraktion
    - Kann Cluster bilden
  - Nachteile:
    - Kein Feedback möglich

Textklassifikation: Beispiel in Perl

- Zur Demonstration folgt eine beispielhafte Implementierung der Vektorraummethode als Perl-Skript
- Das Skript ist dabei keineswegs als vollständig und/oder funktionstüchtig anzusehen
- Es demonstriert lediglich wie einfach ein simples Vektorraummodell umsetzbar ist