ulm university universität ulm

HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN MÜNCHEN

Informationsverwaltung als selbst-organisierendes und kontext-basiertes System

Kerstin Schmidt, Competence Center Wirtschaftsinformatik, Hochschule München

Prof. Dr. Peter Mandl, Competence Center Wirtschaftsinformatik, Hochschule München

Prof. Dr. Michael Weber, Institut für Medieninformatik, Universität Ulm

Agenda

1. **Einführung**
2. **Dokumentendarstellung als Vektorraummodell**
3. **Analyse durch Clusteringverfahren**
4. **Eigene Erweiterung des Vektorraummodells**
5. **Erste Umsetzungen und Beispiele**
6. **Aktuelle Problemstellungen und Ziele**

Agenda

1. **Einführung**
2. **Dokumentendarstellung als Vektorraummodell**
3. **Analyse durch Clusteringverfahren**
4. **Eigene Erweiterung des Vektorraummodells**
5. **Erste Umsetzungen und Beispiele**
6. **Aktuelle Problemstellungen und Ziele**

Einführung (1)

- **„Wissensorganisation“:** nahezu jeder Computer-Nutzer ist bereits im Kleinen mit der Thematik konfrontiert
- **„Wissen und Information“:** abgelegt in einer Vielzahl von Dateien und Dokumenten
- **„Wissensverwaltung“:** strukturierte Ablage oder Mut zum Chaos?

Einführung (2)

- **Basis:** Verschiedene Projekte zum Thema „Workflow- und Dokumenten-Management-Systeme“ (DMS)
- Immer wiederkehrende (Nutzer-)Kritik an DMS:
  - Verwaltungsarbeit zu aufwändig
  - Manuelle Vergabe von Kategorien, Stichwörtern zu zeitraubend, störend und wird daher schlecht/wenig umgesetzt
- **Daher: Forschungs- und Promotionsprojekt**
  - Entwicklung eines Systems, in dem „Dokumente sich selbst einander zuordnen“, um den Nutzer so wenig wie möglich mit manueller Verwaltungsarbeit zu belasten

Einführung (3)

- Grundgedanke: Automatisierte Suche nach „ähnlichen Dokumenten“
- Was heißt „Ähnlichkeit“?
  - Ähnlichkeit basierend auf dem Inhalt
  - Ähnlichkeit basierend auf Autorenkreis
  - ...
- Bekannte Lösung aus dem Information Retrieval: „Vektorraummodell“ kombiniert mit Clusterverfahren

Agenda

1. **Einführung**
2. **Dokumentendarstellung als Vektorraummodell**
3. **Analyse durch Clusteringverfahren**
4. **Eigene Erweiterung des Vektorraummodells**
5. **Erste Umsetzungen und Beispiele**
6. **Aktuelle Problemstellungen und Ziele**

Dokumentendarstellung im Vektorraummodell

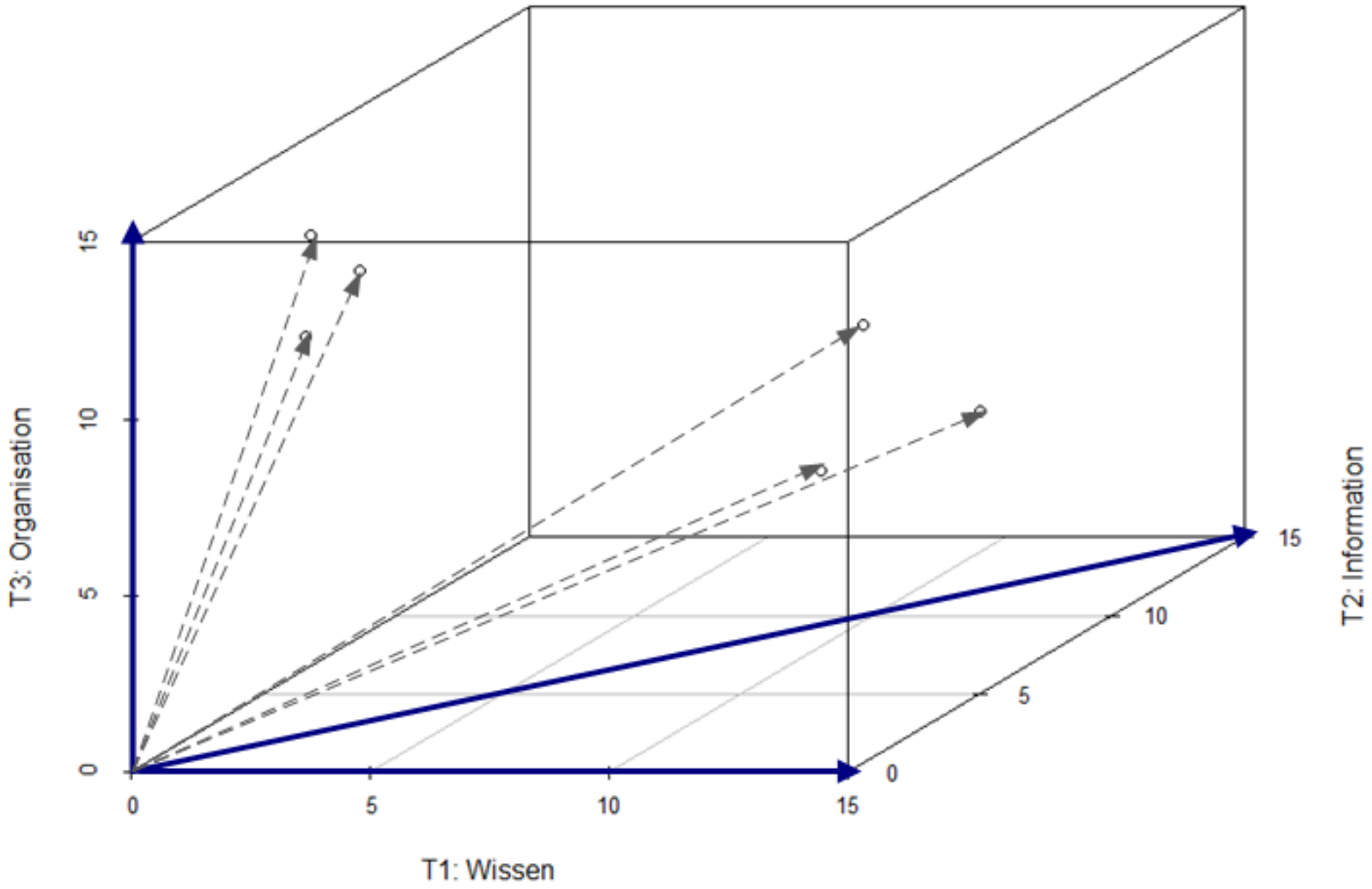
- Bewährter Ansatz aus dem Information Retrieval: Darstellung eines Texts als Vektor
  - Jeder Text/jedes Dokument wird durch einen Vektor repräsentiert (ausschlaggebend ist dabei allein der Inhalt des Textes)
  - Jedes Wort (innerhalb eines Dokumentenpools) stellt eine eigene Dimension des Vektors dar
  - Ähnliche Vektoren sind ähnliche Dokumente

Vektorraummodell: Vereinfachtes Beispiel

- Verschiedene Texte mit den häufigsten Begriffen: „*Wissen*“, „*Organisation*“, „*Information*“

	Wissen	Information	Organisation
Dok 1	15	5	5
Dok 2	10	8	8
Dok 3	12	6	10
Dok 4	2	3	11
Dok 5	1	5	13
Dok 6	2	5	12

Vektorraummodell: Vereinfachtes Beispiel



Optimierung durch Gewichtung der Begriffe

- Analyse der Häufigkeiten von Begriffen kann optimiert werden durch Hinzunahme von Gewichtungen
- Gewichtung durch „TF-IDF-Verfahren“
  - Term wird gewichtet durch seine „Term Frequenz (TF)“ und die „invertierte Dokumenten-Frequenz (IDF)“
  - TF: Wie häufig kommt ein Begriff in einem einzelnen Dokument vor?
  - IDF: Wie viele Texte in eines Dokumentenpools (Korpus) enthalten diesen Begriff?
  - Dokumentenvektor zeigt nicht mehr die absoluten Häufigkeiten der einzelnen Begriffe, sondern die gewichteten Häufigkeiten

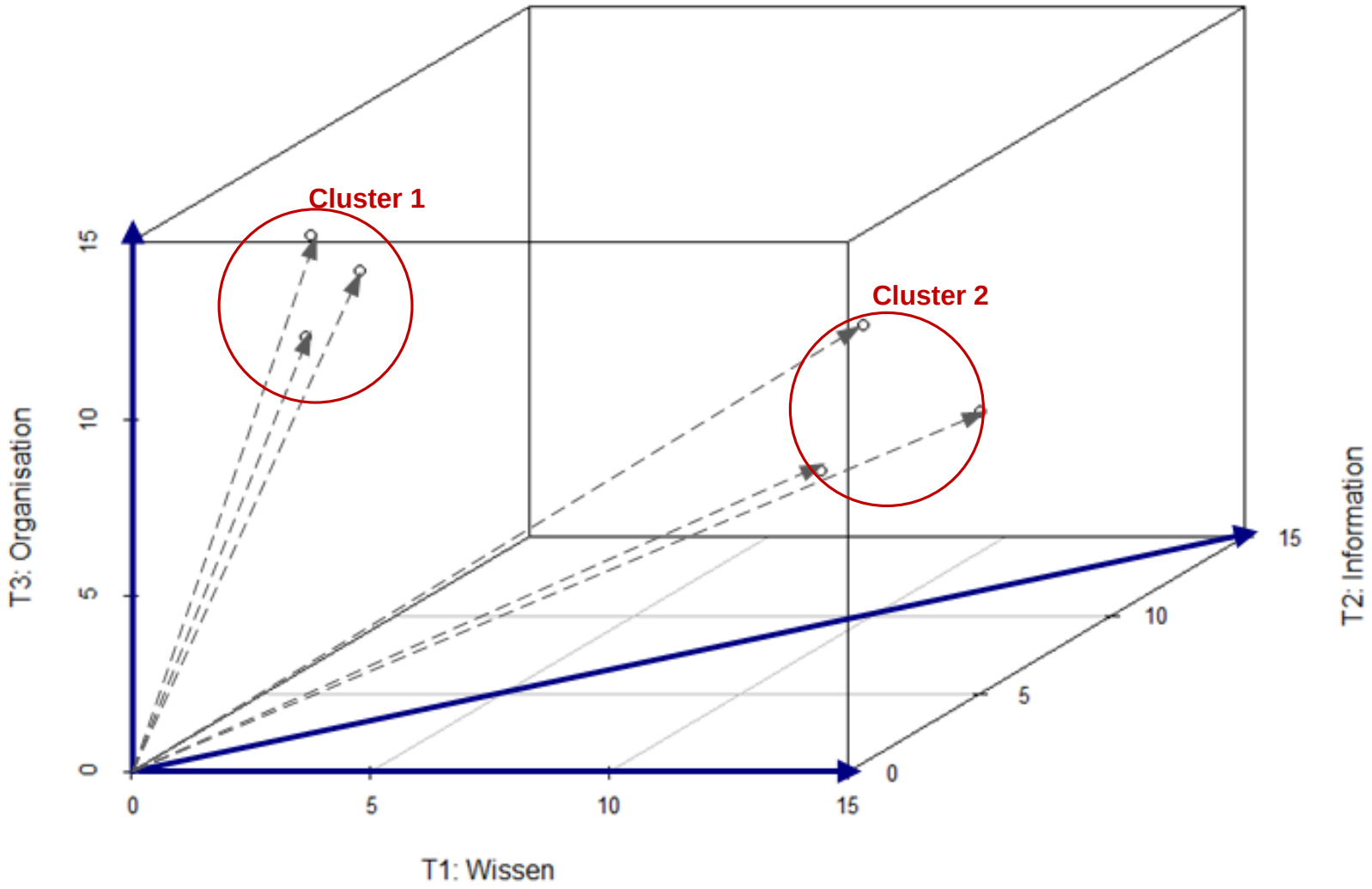
Agenda

1. **Einführung**
2. **Dokumentendarstellung als Vektorraummodell**
3. **Analyse durch Clusteringverfahren**
4. **Eigene Erweiterung des Vektorraummodells**
5. **Erste Umsetzungen und Beispiele**
6. **Aktuelle Problemstellungen und Ziele**

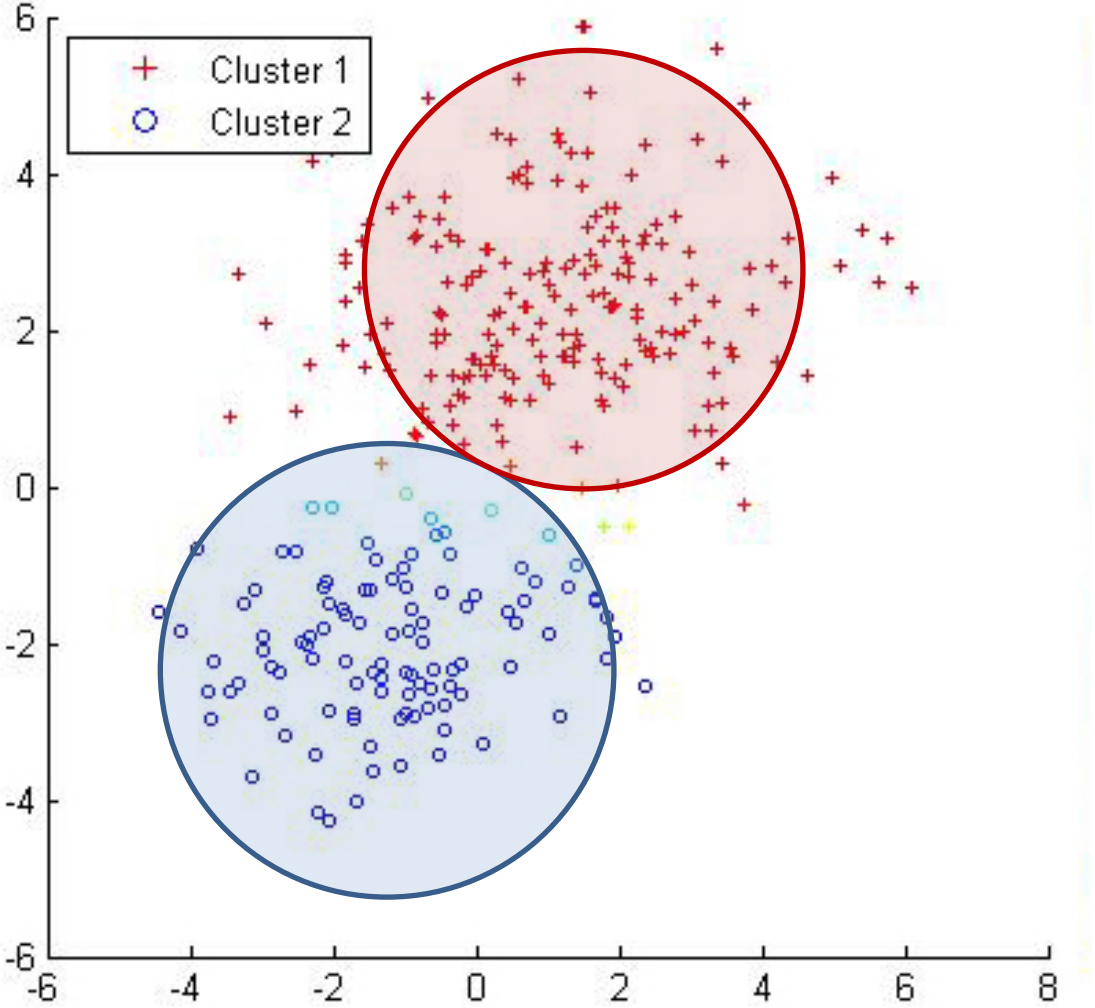
Analyse durch Cluster-Algorithmen

- Suche nach „ähnlichen Vektoren“ erfolgt unter Verwendung von Cluster-Analysen
- Ziel: Automatisches Finden von vergleichbaren Objekt-Gruppierungen („Clustern“)
- Elemente innerhalb eines Clusters verfügen über möglichst ähnliche Eigenschaften und unterscheiden sich so stark wie möglich von Elementen anderer Cluster
- Wichtigste Unterscheidung: Harte oder weiche Clustergrenzen?

Cluster-Untersuchungen im Vektorraum



Harte oder weiche Clustergrenzen?



Agenda

1. **Einführung**
2. **Dokumentendarstellung als Vektorraummodell**
3. **Analyse durch Clusteringverfahren**
4. **Eigene Erweiterung des Vektorraummodells**
5. **Erste Umsetzungen und Beispiele**
6. **Aktuelle Problemstellungen und Ziele**

Erweiterung des Vektorraummodels

- **Klassisches Vektorraummodell**:
  Repräsentation eines Dokuments allein aufgrund der Eigenschaft „Textinhalt“

- **Erweiterung im Rahmen dieses Projekts**:
  „Raumerweiterung“ durch neue Dimensionen, basierend auf zusätzlichen Eigenschaften
  Ziel: Zusammenhänge abbilden, die sich mittels des Inhalts allein nicht darstellen lassen
  - Explizite Eigenschaften (Stichwörter, Dateiname, …)
  - Implizite Eigenschaften (Umgang des Benutzers mit diesen Dateien)

Implizite Dokumenteigenschaften

- Implizite Dokumenteigenschaften: Wie geht der Benutzer mit Dokumenten um?
  - Historie der „Bewegungen“ des Dokuments (Kopieren, Verschieben, Umbenennen)
  - Suchabfragen, bei denen das Dokument geöffnet wurde (entspricht einer Art „unbewussten Stichwortvergabe“)
  - Kontext, in dem das Dokument erstellt wurde (Rollenkontext, sowie – wenn möglich – lokaler Kontext)
- Versuch, „unbewusste Assoziationen“ in den Eigenschaftsraum des Dokuments mit aufzunehmen

Agenda

1. **Einführung**
2. **Dokumentendarstellung als Vektorraummodell**
3. **Analyse durch Clusteringverfahren**
4. **Eigene Erweiterung des Vektorraummodells**
5. **Erste Umsetzungen und Beispiele**
6. **Aktuelle Problemstellungen und Ziele**

Einsatz standardisierter Testkollektionen

- Erste Umsetzungen unter Verwendung von **„R“** (statistische Programmiersprache) und der Verwendung von **„fuzzy“ Cluster-Algorithmen** (also: weiche Clustergrenzen)
- Tests mit standardisierten **Testkollektionen** aus dem Informationen Retrieval
- Umfangreiche Datensammlungen bestehend aus **Texten** und **bekannten Zuordnungen** zu Kategorien
- **Beispiel:** „Reuters-21578“
  - 21578 Texte der Reuters Nachrichtenagentur
  - 135 Kategorien (pro Text: Bekannte Zuordnung zu mindestens einer dieser Kategorien)

Reuters-21578: Häufigste Kategorien

Kategorie	Vorkommen
Earn	3987
Acq	2448
Money-Fx	801
Grain	628
Crude	634
Trade	551
Interest	513
Ship	305
Wheat	306
Corn	254

Beispiel eines Reuters-Dokuments

```
<?xml version="1.0"?>
<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="12550"
NEWID="368">
<DATE>2-MAR-1987 08:25:42.14</DATE>
<TOPICS>
  <D>crude</D>
  <D>ship</D>
</TOPICS>
<PLACES>
    <D>usa</D>
 </PLACES>
  <PEOPLE/>
  <ORGS/>
  <EXCHANGES/>
  <COMPANIES/>
<TEXT>
    <TITLE>PHILADELPHIA PORT CLOSED BY TANKER CRASH</TITLE>
    <BODY>The port of Philadelphia was closed when a Cypriot oil tanker, Seapride
        II, ran aground afterhitting a 200-foot tower supporting power lines
        across theriver, a Coast Guard spokesman said.  He said there was no
        oil spill but the ship is lodged onrocks opposite the Hope Creek nuclear
        power plant in NewJersey.  He said the port would be closed until today
        when theyhoped to refloat the ship on the high tide. […]
    </BODY>
  </TEXT>
</REUTERS>
```

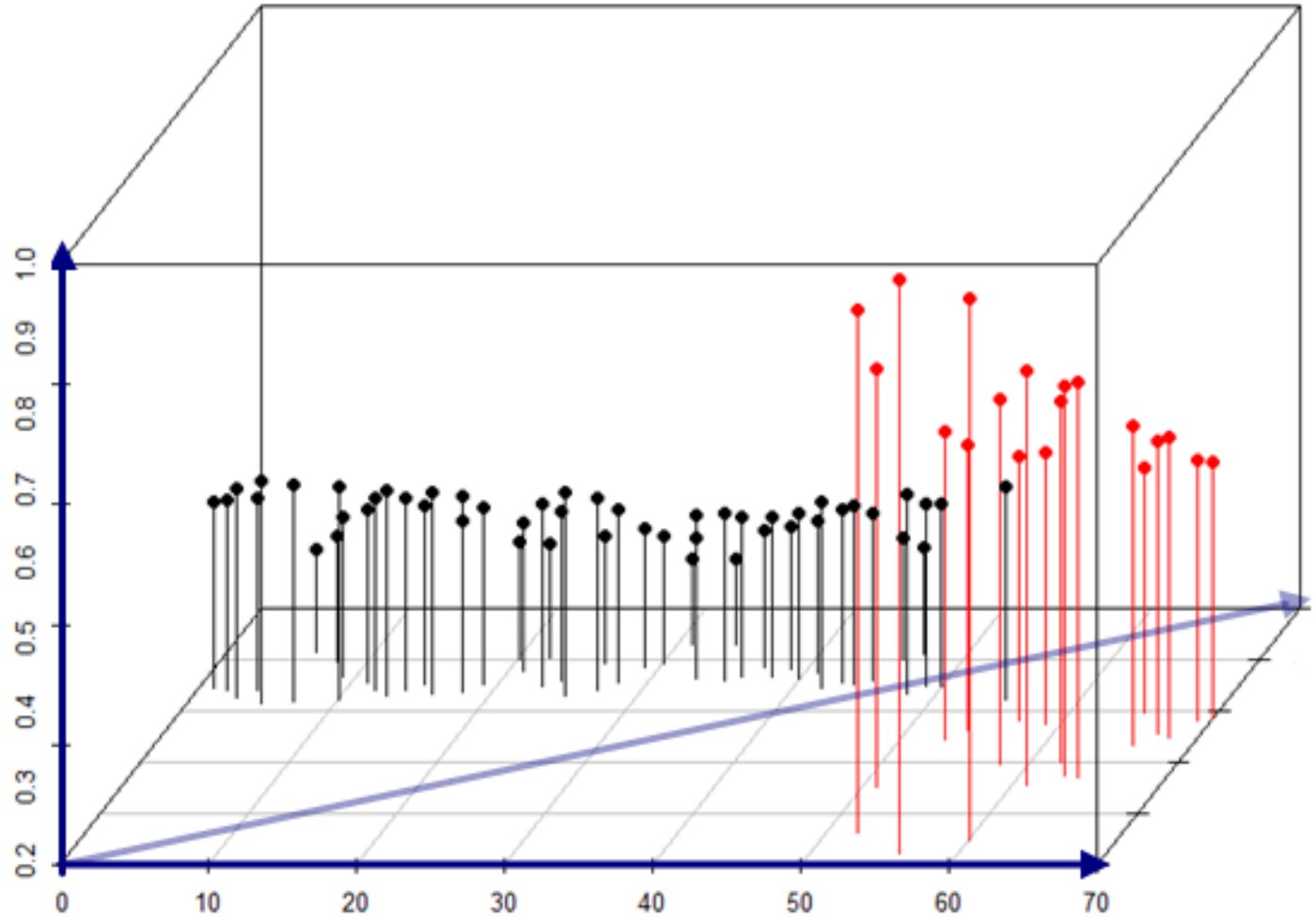
Beispiel und erste Ergebnisse (1)

- **50** Reuters-Dokumente der **Kategorie „crude**
- **20** Reuters-Dokumente der **Kategorie „acq“**
- Text der Dokumente wurde vorverarbeitet
  - Entfernung von Stopwörtern („and“, „the“, …)
  - Bereinigung von Wortendungen
  - Häufigkeiten wurden gewichtet und normalisiert
- Ergebnis: Vektorraum mit **1516 Dimensionen**
- Ergänzung um **nur drei zusätzliche**, externe Eigenschaften, um den erweiterten Vektorraum zu simulieren

Beispiel und erste Ergebnisse (2)

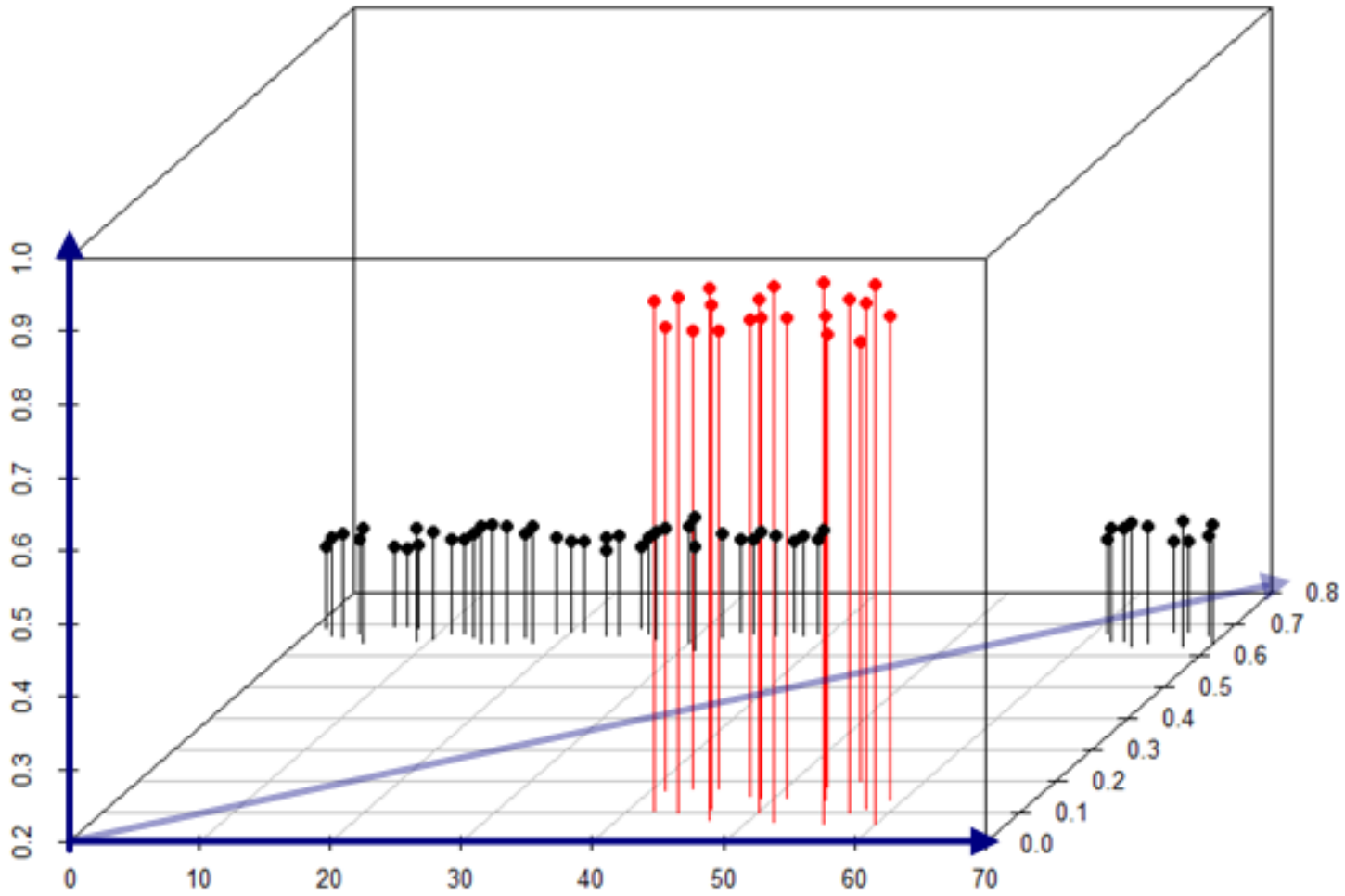
<table>
<tr><th>Dok-ID</th><th>Reuters-Kategorie</th><th colspan="3">Zusätzliche Eigenschaften</th></tr>
<tr><td>Dok-ID 0-10</td><td rowspan="5">Kategorie "Crude"</td><td rowspan="3">Ordner „toRead“</td><td rowspan="3"></td><td rowspan="3"></td></tr>
<tr><td>Dok-ID 11-20</td></tr>
<tr><td>Dok-ID 21-30</td></tr>
<tr><td>Dok-ID 31-40</td><td></td><td rowspan="2">Ordner „article“</td><td rowspan="2">Autor „Schmidt“</td></tr>
<tr><td>Dok-ID 41-50</td><td rowspan="3">Ordner „toRead“</td></tr>
<tr><td>Dok-ID 51-60</td><td rowspan="2">Kategorie "acq"</td><td rowspan="2"></td></tr>
<tr><td>Dok-ID61-70</td></tr>
</table>

Ergebnis nur auf Basis des Inhalts



Hinweis zum Diagramm:
X-Achse: Index der Dok-IDs; Y- und Z-Achse: Zugehörigkeit zu Cluster 1 und 2

Ergebnis inkl. zusätzlicher Eigenschaften



Hinweis zum Diagramm:
X-Achse: Index der Dok-IDs; Y- und Z-Achse: Zugehörigkeit zu Cluster 1 und 2

Agenda

1. **Einführung**
2. **Dokumentendarstellung als Vektorraummodell**
3. **Analyse durch Clusteringverfahren**
4. **Eigene Erweiterung des Vektorraummodells**
5. **Erste Umsetzungen und Beispiele**
6. **Aktuelle Problemstellungen und Ziele**

Aktuelle Problemstellungen

- Gewichtung: Text - zusätzliche Eigenschaften
- **Dokumentenstruktur:**
  - Heterogene Struktur mit verschiedensten Themen (hohe Zahl an Dimensionen)
  - Homogene Struktur mit relativ gleichartiger Thematik (geringere Anzahl an Dimensionen)
- **Benutzertypen:**
  - Aktiver Umgang mit Dokumenten (zahlreiche zusätzliche Dimensionen)
  - Passiver, eher „sammelnder“ Umgang (wenige Dimensionen)

Ziel

- Entdecken von Zusammenhängen zwischen Dokumentenstruktur, Benutzertyp und verschiedenen Cluster-Algorithmen

- Idealvorstellung: Jeweils passender Clusteralgorithmus pro Kombination aus Dokumentenstruktur und Benutzertyp

- Ziel: Optimal angepasste Software-Umsetzung je nach Benutzertyp und Dokumentenpool

Vielen Dank für Ihr Interesse!

?

Fragen und Antworten

!