

# Clustering von Patent-Dokumenten

Magisterarbeit IIM  
Joachim Pfister

HIER 2005 –  
Hildesheimer Evaluierungs- und Retrieval Workshop  
20.07.2005, Universität Hildesheim

## Rahmen

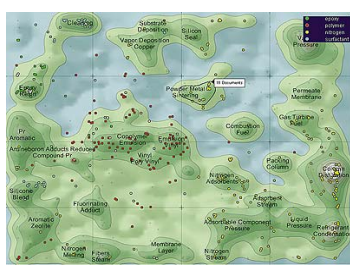
- o **Titel der Magisterarbeit:**  
„Analyse und Einsatzpotentiale von Clustering-Verfahren zum Retrieval von Patent-Dokumenten“
- o Betreuer: Prof. Womser-Hacker, Dr. Mandl
- o In Kooperation mit dem FIZ-Karlsruhe (Dr. Schwantner)
- o Bearbeitungszeitraum: März – November 2004
- o **Zielsetzung:**  
Können automatisch erzeugte Cluster eine Hilfe für den Benutzer darstellen beim Navigieren in einer Dokumentenmenge als Antwort auf eine Suchanfrage?

## Clustering

- o **clustern** = Objekte in Gruppen einteilen, wobei
  - Objekte innerhalb einer Gruppe möglichst ähnlich
  - Gruppen möglichst unterschiedlich, d.h. separiert
- o Grundlage: Cluster-Hypothese von van Rijsbergen  
„[...] closely associated documents tend to be relevant to the same requests.“ (van Rijsbergen 1979, 30)
- o Arten des Clusters bei Dokumentensammlungen:
  - Pre-Retrieval Clustering
  - Post-Retrieval Clustering

## Post-Retrieval Clustering

### Visualisieren von Suchergebnissen



Darstellung einer Clustering-Lösung durch ThemeScape

Quelle:  
<http://www.researchinformation.info/rjanfeb04patent1.html>

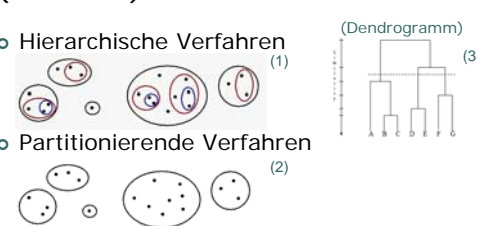
## Post-Retrieval Clustering

### Probleme bzw. Herausforderungen

- o **Automatische Benennung von Clustern**
  - Strategien zum Ermitteln der Benennungen (z.B. via TF/IDF o.Ä.)
  - Nutzer können sich durch falsche Benennungen getäuscht fühlen
- o **Nachvollziehbarkeit der erzeugten Clustering-Lösungen?**
  - Ansprüche der Nutzer zu hoch für (bisher) technisch umgesetzte Lösungen?
  - Verschiedene Menschen = verschiedene Einteilungsmöglichkeiten (dabei: geringe Übereinstimmung), vgl. Macskassy et al. (1998)
- o „Clustering did not appear to be preferable to ranked lists especially as it also represented overheads in both computing time and resources involved in creation of the clusters...“ (Kural et al. 2001, 596)

## Arten von Clustering- / Fusionierungsverfahren (Auswahl)

- o **Hierarchische Verfahren** (1)
- o **Partitionierende Verfahren** (2)
- o **Probabilistische Verfahren**  
→ Zugehörigkeitswahrscheinlichkeit einer Instanz zu einer Klasse

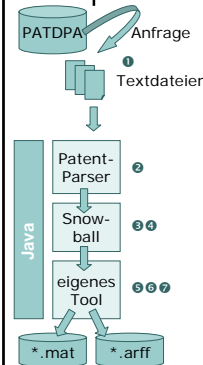


6 1, 2: (Chojnacki 2003), 3: (Jain et al. 1999)

## Auswahl der Clustering-Verfahren für die Experimente

Verfahren	SW	Bemerkungen
bi-secting K-Means (partitionierendes Verfahren)	CLUTO (Clustering Toolkit)	<ul style="list-style-type: none"> <li>Vergleichsuntersuchung: Zhao und Karypis (2003)</li> <li>Clusteranzahl durch manuelle Vorgabe</li> </ul>
Shortest Nearest Neighbor	SNN	<ul style="list-style-type: none"> <li>Ertöz et al. (2003)</li> <li>Clusteranzahl automatisch</li> </ul>
probabilist. Verf.	Autoclass-C	<ul style="list-style-type: none"> <li>Clusteranzahl automatisch</li> </ul>
probabilist. Verf. (Expectation Maximization)	WEKA	Erwies sich in den Vorab-Versuchen als ungeeignet (Meist ein großer Cluster, in dem fast sämtliche Objekte lagen).
Pseudo-Lösung	---	Cluster = gleiche Main-IPC

## Datenaufbereitung



1. Patentedokumente aus der DB „PATDPA“ im Format „brief“ ausgeben → Textdatei
2. Extrahieren der für Experimente relevanten Inhalte (TI, MCLM, AB, Main-IPC)
3. Stoppwörter entfernen
4. Stemming
5. Term-Gewichtung nach Okapi-BM-25
6. Dokument/Term-Matrix erstellen
7. Erstellen der Quelldateien für die versch. Clustering-Tools

## Datengrundlage

- Keine „richtigen“ (= in der Praxis anfallenden) Anfragen an die DB PATDPA z.B. aus Log-Dateien extrahiert (Datenschutz), sondern selbst erdachte und möglichste „praxisnahe“ Anfragen.
- Einschränkung der Anfragen auf eine Hauptklasse der IPC (G06F017), um thematisch kohärentere Cluster zu erzeugen.
- Mindestanzahl von Termen (nach Stoppwort-Elimination und Stemming) von 5 Termen pro Dokument
- Suchanfragen für Experimente liefern mindestens 80 Dokumente als Ergebnis zurück

9

## Ziele der Experimente

- Annahme 1: *Das Entfernen von Patentfamilien-Doppeln in den Ausgangsdaten führt zu einer besseren Clusterqualität.*
- Annahme 2: *Ein Verfahren zur Erzeugung von Clustering-Lösungen sticht mit qualitativ hochwertigen Lösungen deutlich hervor.*
- Annahme 3: *Die Gruppierung von Patentedokumenten mittels der IPC-Klassen ist per se ideal.*

10

## Evaluierungsansätze

### Ansätze zur Bewertung

- objektiv (vgl. Jain und Dubes, 1988)
  - Extern:** Vergleich mit existierender Struktur
  - Intern:** Stellt die ermittelte Struktur für die Ausgangsdaten eine passende Beschreibung dar (ohne Rückgriff auf externe Informationen)?
  - Relativ:** Vergleich von zwei Ergebnissen (evtl. mit weiteren Maßzahlen)
- subjektiv (Nutzerinteresse)
  - „Cluster Usability“ (vgl. Stein et al., 2003) mittels Relevanzurteilen der Nutzer

11

## Cluster Usability - Vorgehen

### Relevanzbewertung durch Juroren

- 12 Studierende (keine Experten)
- Aufteilung in 3 Gruppen
- 3 Clustering-Verfahren, Gruppe A mit Pseudo-Lsg.

<b>A</b> Juroren A1–A3	bild? (S) verarbeit? AND G06F017/ICM medizin? AND G06F017/ICM bild_verarbeit_ipc_md medizin_ipc_md
<b>B</b> Juroren B1–B3	datenuebertragung? AND G06F017/ICM server? AND client? AND G06F017/ICM digital? AND bild? AND G06F017/ICM
<b>C</b> Juroren C1–C3	brows? AND G06F017/ICM multimedia? AND G06F017/ICM navig? AND G06F017/ICM

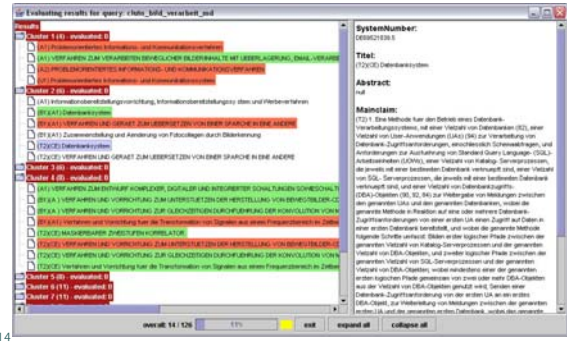
12

## Cluster Usability - Durchführung

- Einbettung der Juroren in eine Recherche-Situation (in Aufgabenbeschreibung)
- Relevanzbewertung mittels ClustEv:** Stehen die in einem Cluster zusammengefassten Dok. in einer inhaltlichen Beziehung? (Grobkonzept o.Ä.) → „Passt“/„Passt nicht“-Entscheidung pro Dok.
- Papierfragebogen:**
  - Schulnotenskala pro Anfrage
  - Clusteranzahl passend? (pro Anfrage)
  - Gesamteindruck aller bewerteten Anfragen
  - Freie Bemerkungen

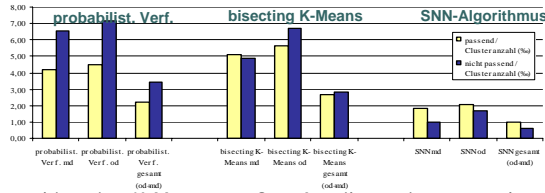
13

## ClustEv – Tool zur Relevanzbewertung durch die Juroren



14

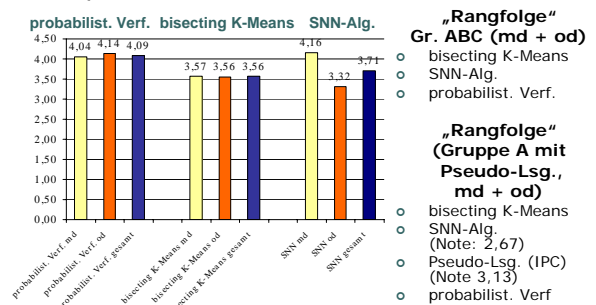
## Ergebnisse Relevanzbewertung auf Dokumentenebene



- bisecting K-Means:** größter Anteil an mit „passend“ bewerteten Dok.
- probabilistisches Verf.:** mehrheitlich mit „nicht passend“ bewertet
- SNN-Algorithmus:** Anteil mit Bewertung „passend“ überwiegt, jedoch schwächer ausgeprägt als bei bisecting K-Means.

15

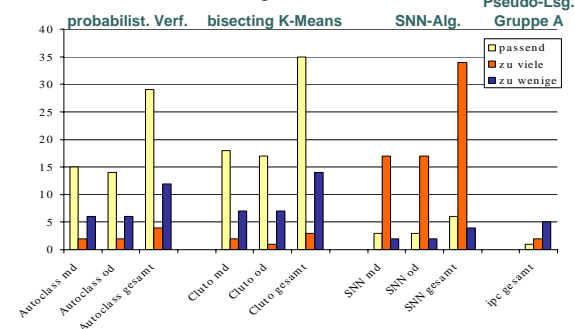
## Beurteilung nach Schulnoten



- „Rangfolge“ Gr. ABC (md + od)**
  - bisecting K-Means
  - SNN-Alg.
  - probabilist. Verf.
- „Rangfolge“ (Gruppe A mit Pseudo-Lsg., md + od)**
  - bisecting K-Means
  - SNN-Alg. (Note: 2,67)
  - Pseudo-Lsg. (IPC) (Note 3,13)
  - probabilist. Verf

16

## Anzahl der erzeugten Cluster okay?



17

## Ergebnisse (I.)

### Annahme 1:

Das Entfernen von Patentfamilien-Doppeln in den Ausgangsdaten führt zu einer besseren Clusterqualität.

- Spielte keine Rolle bei der Bewertung durch die Juroren (Relevanzbewertung, Schulnoten, Kommentare)
- Ergebnis:** Entfernen führt zu keiner Qualitätsverbesserung.

18

## Ergebnisse (II.)

### Annahme 2:

Ein Verfahren zur Erzeugung von Clustering-Lösungen sticht mit qualitativ hochwertigen Lösungen deutlich hervor.

- Bewertung nach Schulnoten liegen in einem Spektrum von 3,56 bis 4,16 dicht beieinander.
- Große Abhängigkeit von der gewählten Anfrage  
→ Anfrage wichtiger als das getestete Clustering-Verfahren?  
(Gruppen A bewertete z.B. häufig mit „gut“, während andere Gruppen generell schlechter bewerteten)
- **Ergebnis:**  
Geringfügiger Vorsprung für das bisecting K-Means Verfahren (Schulnoten, Relevanzbewertung)

19

## Ergebnisse (III.)

### Annahme 3:

Die Gruppierung von Patentedokumenten mittels der IPC-Klassen ist per se ideal.

- Nur von Gruppe A getestet (alle 4 Juroren)
- Schlechtes Abschneiden bei der Bewertung mittels Schulnoten (Note: 3,13; Platz 3) und in den Kommentaren der Juroren.
- **Ergebnis:**  
Kaum vergleichbar mit den Gesamturteilen, jedoch tendenziell mit eher schlechteren Bewertungen versehen.

20

## Einflüsse auf die Ergebnisse haben...

- **Datengrundlage**
  - Auswahl der Anfragen
  - Experimente wurden nur mit Teilen von Patentedokumenten durchgeführt.  
→ mögl. Erweiterung:  
Clustering mit Volltexten
- **Art und Weise der Datenaufbereitung**
  - Stoppwort-Elimination (Umfang und Inhalt der Liste)  
→ mögl. Erweiterung:  
Anfragespezifische Stoppwortliste
  - Stemming-Algorithmus (dessen Fähigkeiten z.B. bei der Kompositzerlegung)
  - Schema zur Termgewichtung (TF/IDF, Okapi-BM 25)
  - Mindestanzahl an 5 Termen pro Dok.
- **Juroren** (Laien, mangelnde Fachkenntnisse)

21

## Danke!

Fragen?  
Anregungen?  
Kommentare?

22

## Literatur (I.)

- Chojnacki, Michael: Text-Clustering. Foliensatz der Präsentation im Rahmen des Blockseminars „Invisible Web“. Universität Duisburg-Essen, 2003
- ERTÖZ, Levent ; STEINBACH, Michael ; KUMAR, Vipin: *Finding Topics in Collections of Documents: A Shared Nearest Neighbor Approach*. S. 83–103. In: WU, Weili (Hrsg.); XIONG, Hui (Hrsg.); SHEKHAR, Shashi (Hrsg.): *Clustering and Information Retrieval*. Dordrecht : Kluwer Academic Publishers, 2003
- JAIN, A. K. ; DUBES, R. C.: *Algorithms for Clustering Data*. Upper Saddle River, NJ : Prentice-Hall, 1988
- Jain, A.K., Murty M.N., and Flynn P.J.: Data Clustering: A Review, ACM Computing Surveys, Vol 31, No. 3, 264-323. <http://citeseer.ist.psu.edu/jain99data.html>, 1999.
- KURAL, Yasemin ; ROBERTSON, Steve ; JONES, Susan: Deciphering cluster representations. In: *Information Processing and Management* 37 (2001), Nr. 4
- MACSKASSY, Sofus A. ; BANERJEE, Arunava ; DAVISON, Brian D. ; HIRSH, Haym: Human Performance on Clustering Web Pages / Department of Computer Science Rutgers, The State University of New Jersey. URL <http://www.cs.rutgers.edu/pub/technical-reports/dcs-tr-355.ps.Z> – Zugriffsdatum:06.10.2004, 11:23 MEZ, 1998 (DCS-TR-355). – Forschungsbericht

23

## Literatur (II.)

- PFISTER, Joachim: Analyse und Einsatzpotentiale von Clustering-Verfahren zum Retrieval von Patent-Dokumenten. Universität Hildesheim. Institut für Angewandte Sprachwissenschaft. Magisterarbeit. Betreuer: Prof. Dr. Womser-Hacker, Dr. Mandl, 2004
- RIJSBERGEN, C. J. van: *Information Retrieval*. Second Edition. London : Butterworths, 1979
- STEIN, Benno ; EISSEN, Sven Meyer zu ; WISSBROCK, Frank: On Cluster Validity and the Information Need of Users. Benalmadena, Spain : ACTA Press, September 2003, S. 216–221
- ZHAO, Ying ; KARVPIS, George: Hierarchical Clustering Algorithms for Document Datasets / University of Minnesota. Department of Computer Science and Engineering. URL [https://www.cs.umn.edu/tech-reports\\_upload/tr2003/03-027.pdf](https://www.cs.umn.edu/tech-reports_upload/tr2003/03-027.pdf) – Zugriffsdatum: 08.10.2004, 13:01 Uhr MEZ, 2003 (#03-027). – Forschungsbericht

24