Dokumenten-Clustering

Norbert Fuhr

Dokumenten-Clustering

(Dokumenten-)Cluster: Menge von ähnlichen Dokumenten
Ausgangspunkt „Cluster-Hypothese“:
die Ähnlichkeit der relevanten Dokumente untereinander und der irrelevanten Dokumente untereinander ist größer als die zwischen anderen (zufälligen) Teilmengen der Dokumentenkollektion
(experimentell nachgewiesen von Rijsbergen und Sparck Jones 1972)
Ziel des Clustering:
Bestimmung dieser Cluster unabhängig von Fragen (schon beim Aufbau der Kollektion)

Prinzipielle Vorgehensweise:

1. Festlegung eines Ähnlichkeitsmaßes (z.B. Skalarprodukt oder Cosinus-Maß)
2. Berechnung der Ähnlichkeitmatrix für alle möglichen Dokumentenpaare aus $|\underline{D}|$
3. Berechnung der Cluster
4. Physisch gemeinsame Abspeicherung der Dokumente eines Clusters

Gliederung

Agglomeratives Clustering

1. Wahl eines Schwellenwertes α für die Ähnlichkeit
2. für alle Dokumente:
   füge d_k zu Cluster C_l hinzu falls
   a) single link-Clustering:

$$\min_{d_i \in C_l} sim(d_k, d_i) \geq \alpha$$

   b) complete link-Clustering:

$$\max_{d_i \in C_l} sim(d_k, d_i) \geq \alpha$$

   c) average link-Clustering:

$$\frac{1}{|C_l|} \sum_{d_i \in C_l} sim(d_k, d_i) \geq \alpha$$

3. falls es kein solches Cluster gibt, bildet d_k ein neues Cluster.

Aufwand für Clustering beträgt $O(n^2)$!

Gliederung

Partitionierendes Clustering (k-means)

1. wähle Anzahl k zu bildender Cluster
2. bestimme k "seed"-Dokumente, die hinreichend unterschiedlich sind. Diese bilden jeweils den Kern eines der Cluster $C_1, \ldots, C_k$
3. für alle (übrigen) Dokumente d_i:
   füge d_i zu dem ähnlichsten Cluster hinzu
4. Wähle Zentroiden der resulierenden Cluster als neue "seeds"
5. Wiederhole Schritte 3 und 4, bis die Cluster stabil sind.

Aufwand: $O(kn)$

Gliederung

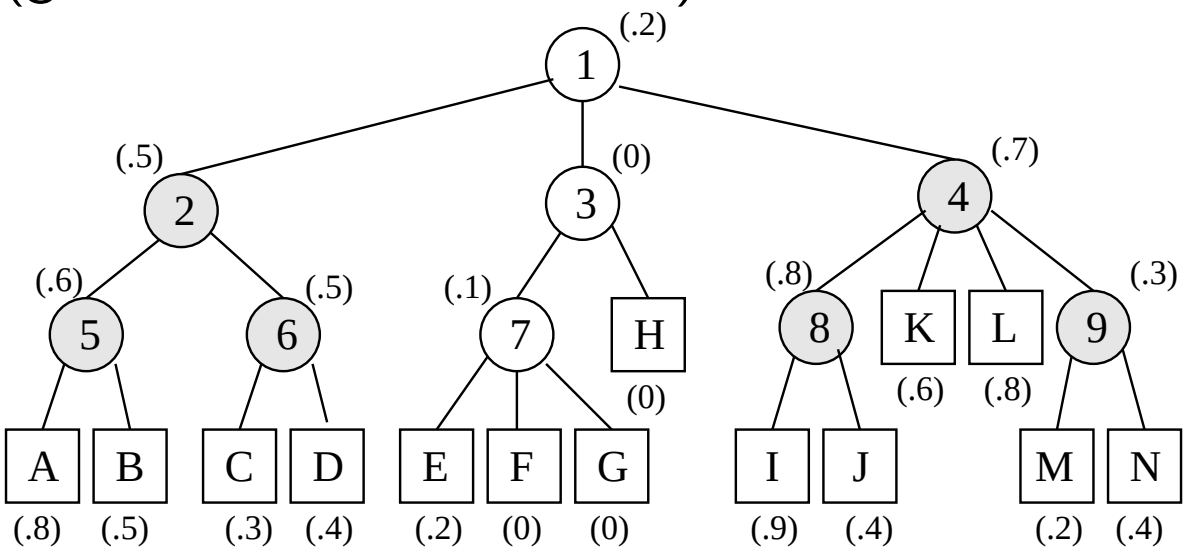
Hierarchisches Clustering

- Bottom up
  - Start: jede Instanz = 1 Cluster
  - In jedem Schritt: vereinige die beiden Cluster mit der kleinsten Instanz
  - Entwurfsentscheidung: Distanz zwischen Clustern
    z.B. als kleinster/größter Abstand zwischen zwei Instanzen, oder als Distanz der Zentroiden.
- Top down
  - Start: alle Instanzen in einem Cluster
  - Aufteilung in zwei Cluster
  - Rekursive Prozessierung jedes erzeugten Clusters
  - sehr effizient

Gliederung

Cluster-Suche

zu jedem Cluster wird ein Zentroid berechnet
(virtuelles Dokument mit minimalem Abstand zu allen Dokumenten des Clusters)
gemeinsame Abspeicherung der Zentroiden
(getrennt von den Clustern)



Retrieval

1. Bestimmung der Zentroiden mit den höchsten Retrievalgewichten
2. Ranking der Dokumente in den zugehörigen Clustern

Beurteilung:

+ Abhängigkeiten zwischen Dokumenten werden berücksichtigt (im Gegensatz zu allen anderen Modellen)
+ weniger I/O als bei normaler Suche
– schlechtere Retrievalqualität
+ es werden andere relevante Dokumente gefunden

Ähnlichkeitssuche und Browsing von Dokumenten

Ähnlichkeitssuche: nur anwendbar, wenn ein relevantes Dokument bekannt
Ziel: Suche nach dazu ähnlichen Dokumenten
(erspart die Formulierung einer Anfrage)

a) über die vorher berechneten Cluster
b) analog zum Vektorraum-Modell
(interpretiere Dokumentvektor als Fragevektor)

Experimentelle Ergebnisse:

- Ähnlichkeitssuche sinnvoll als Ergänzung zu den anderen Retrievalmodellen
  (es werden andere relevante Dokumente gefunden)
- Clustering ermöglicht Browsing
- Vorprozessierung der Cluster nur für Retrieval lohnt nicht

Gliederung

Probabilistisches Clustering

Verallgemeinerung von k-means-clustering auf "unscharfe" Cluster

$C^1, \ldots, C^k$ Cluster

$\mathbf{x} = (x_1, \ldots, x_n)$: Merkmalsvektor eines Dokumentes d mit

$$x_i = \begin{cases} 1, & \text{falls } t_i \in d^T \\ 0, & \text{sonst} \end{cases}$$

Wahrscheinlichkeit, dass Dokument mit Vektor $\mathbf{x}$ zu Cluster C^j gehört: $P(C^j|\mathbf{x})$

Anwendung des Bayes'schen Theorems

$$P(a|b) = \frac{P(a,b)}{P(b)} = \frac{P(b|a) \cdot P(a)}{P(b)}$$

$$\begin{aligned} P(C^j|\mathbf{x}) &= \frac{P(\mathbf{x}|C^j)P(C^j)}{P(\mathbf{x})} \\ &= \frac{P(\mathbf{x}|C^j)P(C^j)}{\sum_{l=1}^{k} P(C^l)P(\mathbf{x}|C^l)} \end{aligned}$$

Unabhängigkeitsannahme:

$$
\begin{aligned}
P(\mathbf{x}|C^j) &= \prod_i P(x_i|C^j) \\
&= \prod_{x_i=1} P(x_i = 1|C^j) \cdot \prod_{x_i=0} P(x_i = 0|C^j)
\end{aligned}
$$

$P(C^j)$ Wahrscheinlichkeit, dass beliebiges Dokument zum Cluster C^j gehört

$q_i^j = P(x_i = 1|C^j)$ Wahrscheinlichkeit, dass Term t_i in einem zufälligen Dokument des Clusters C^j vorkommt

$$
P(\mathbf{x}|C^j) = \prod_{x_i=1} q_i^j \cdot \prod_{x_i=0} (1 - q_i^j)
$$

Parameterschätzung

Cluster sind nicht von vornherein bekannt
$\rightarrow$ Anwendung des EM-Algorithmus' (expectation maximization)

1. E: Berechne die Cluster-Wahrscheinlichkeit für jede Instanz
2. M: Schätze die Parameter basierend auf den Cluster-Wahrscheinlichkeiten

$$
\begin{aligned}
n^j &= \sum_{d_m \in D} P(C^j|\mathbf{x}_m) \\
p^j &= \frac{n^j}{|D|} \\
q_i^j &\approx \frac{1}{n^j} \sum_{d_m \in D} x_{m_i} \cdot P(C^j|\mathbf{x}_m) \\
&\approx \frac{1}{n^j + 1} \left(p^j + \sum_{d_m \in D} x_{m_i} \cdot P(C^j|\mathbf{x}_m) \right)
\end{aligned}
$$

Anwendung

1. wähle Anzahl k zu bildender Cluster
2. bestimme k "seed"-Dokumente, die hinreichend unterschiedlich sind. Diese bilden jeweils den Kern eines der Cluster $C^1, \ldots, C^k$
3. Initialisierung der Parameter: Setze $n^j = 1$ und $p^j = 1/k$. Ferner sei

$$P(C^j|\mathbf{x}_m) = \begin{cases} 1, & \text{falls } d_m \text{ seed von } C^j \\ 0, & \text{sonst} \end{cases}$$

   Berechne daraus initiale Werte für die q_i^j
4. Für alle Dokumente $d_m \in D$: Berechne $P(C^j|\mathbf{x}_m)$ für $j = 1 \ldots, k$
5. Berechne neue Parameter n^j, p^j und q_i^j
6. Wiederhole die letzten beiden Schritte, bis die Cluster stabil sind.

Erweiterung auf nummerische Merkmale

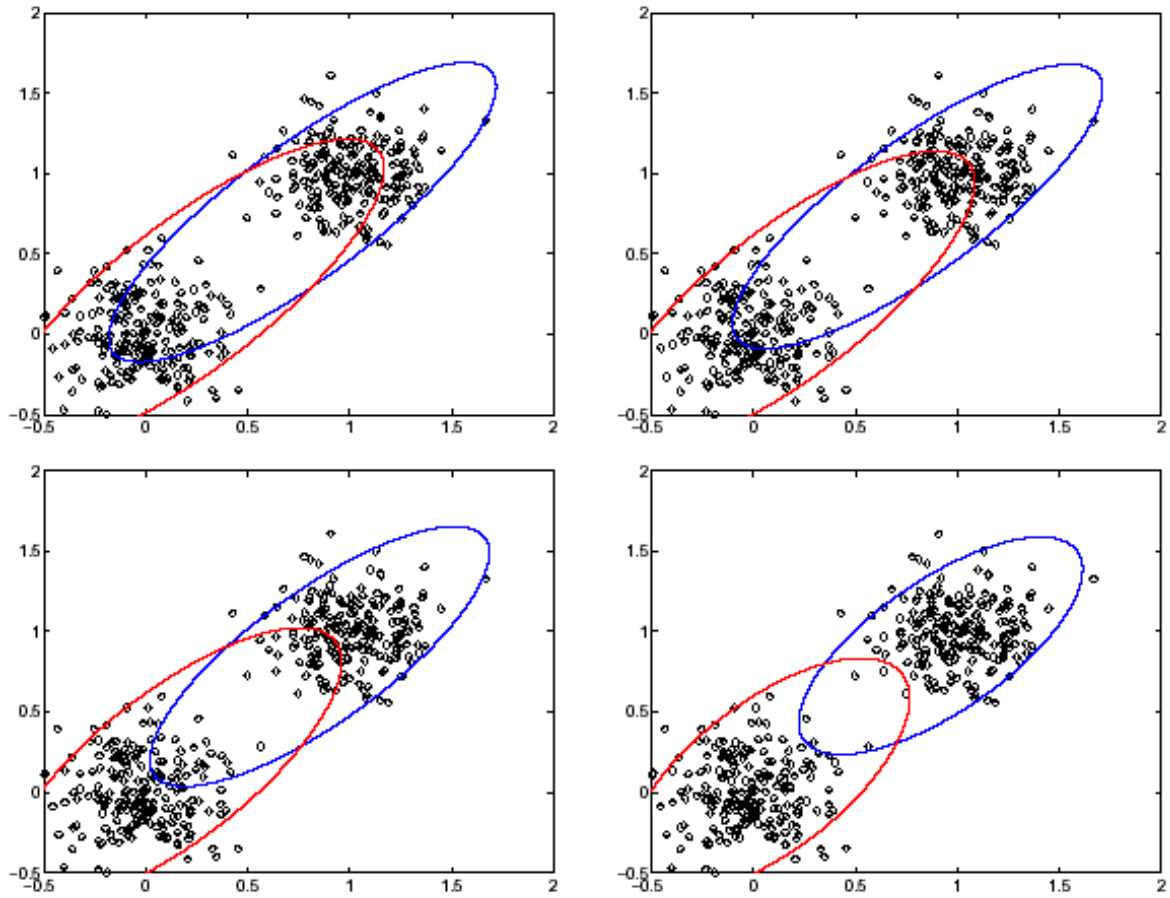
Annahme einer Normalverteilung:

$$P(x_i|C^j) = \frac{1}{\sqrt{1\pi\sigma_i}} e^{\frac{(x_i-\mu_i^j)^2}{2(\sigma_i^j)^2}}$$

erfordert Schätzung von 2 Parametern pro Merkmal und Cluster:

$$\begin{aligned} \mu_i^j &= \frac{1}{n^j} \sum_{d_m \in D} x_{m_i} \cdot P(C^j|\mathbf{x}_m) \\ \sigma_i^j &= \frac{1}{n^j} \sum_{d_m \in D} (x_{m_i} - \mu_i^j)^2 \cdot P(C^j|\mathbf{x}_m) \end{aligned}$$

Beispiel zu probabilistischem Clustering



Gliederung

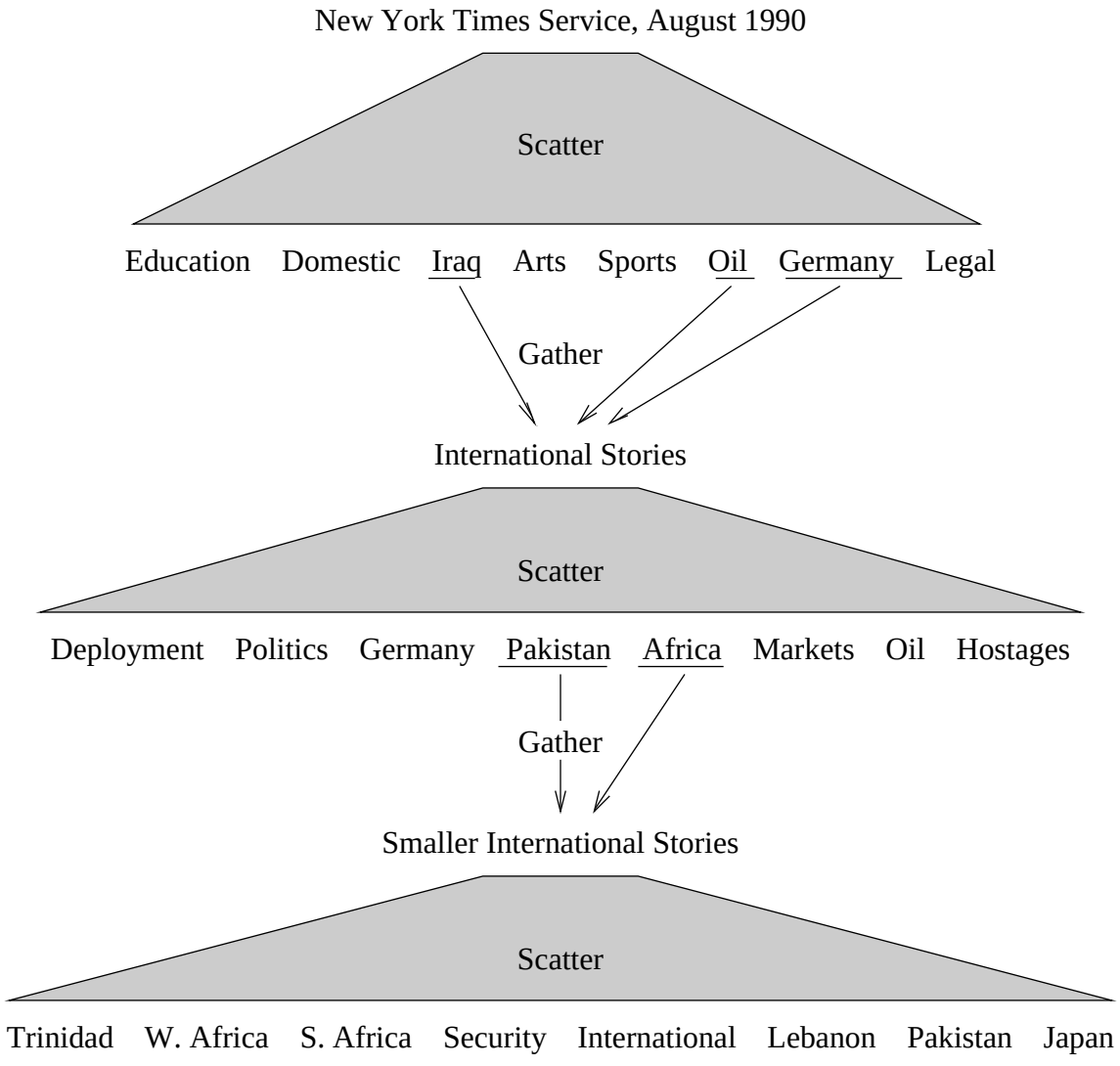
Scatter-Gather-Clustering

Browsing durch eine dynamisch generierte Hierarchie
(basiert auf partitionierendem Clustering)
Scatter = "Zerstreuen"

- Aufteilen der Ausgangsmenge in Gruppen
- partitionierendes Clustering

Gather = "Sammeln"

- Anwender wählt Gruppen aus
- Gruppen werden zusammengefasst
- neue Ausgangsmenge


New York Times Service, August 1990
Scatter
Education
Domestic
Iraq
Arts
Sports
Oil
Germany
Legal
Gather
International Stories
Scatter
Deployment
Politics
Germany
Pakistan
Africa
Markets
Oil
Hostages
Gather
Smaller International Stories
Scatter
Trinidad
W. Africa
S. Africa
Security
International
Lebanon
Pakistan
Japan

Cluster 1 Size: 4 assistant director deputy secretary special affair division administrator management staff pc

- 603252 "Excepted Service; Consolidated Listing of Schedules A, B, and C Exceptions"
- 329912 "Excepted Service; Consolidated Listing of Schedules A, B, and C Exceptions"
- 610814 "5 CFR Part 737"
- 317319 "SES Positions That Were Career Reserved During 1988"

Cluster 2 Size: 187 deposit capital asset insurance risk fail save credit rate market account billion

- 631435 "World Business (A Special Report): Eastern Europe --- The Idea Man: France's Jacques Attali Is the Driving Force Behi
- 658624 "Politics & Policy: CIA Warned In '86 of Entry Of BCCI to U.S. ---- By Peter Truell Staff Reporter of The Wall Street Journ
- 39340 "House, Senate Versions Compared"
- 402897 "Under Fire: World Bank's Conable Runs Into Criticism On Poor Nations' Debt --- Liberals Assail His Refusal To Give M
- 333197 "Federal Reserve Bank Services"

Cluster 3 Size: 217 section information 2 requirement regulation 3 request rule record 5 provision procedure

- 690665 "Security is big business. (balancing security systems and user training to achieve data security)"
- 592791 "Organization; Farm Credit System Financial Assistance Corp."
- 322941 "PART 78_EDUCATION APPEAL BOARD"
- 334160 "12 CFR Parts 7 and 32"
- 334479 "Privacy Act of 1974; Systems of Records"

Cluster 4 Size: 85 investigation allege fraud court lawyer firm prosecutor jury bcci american grand defendant

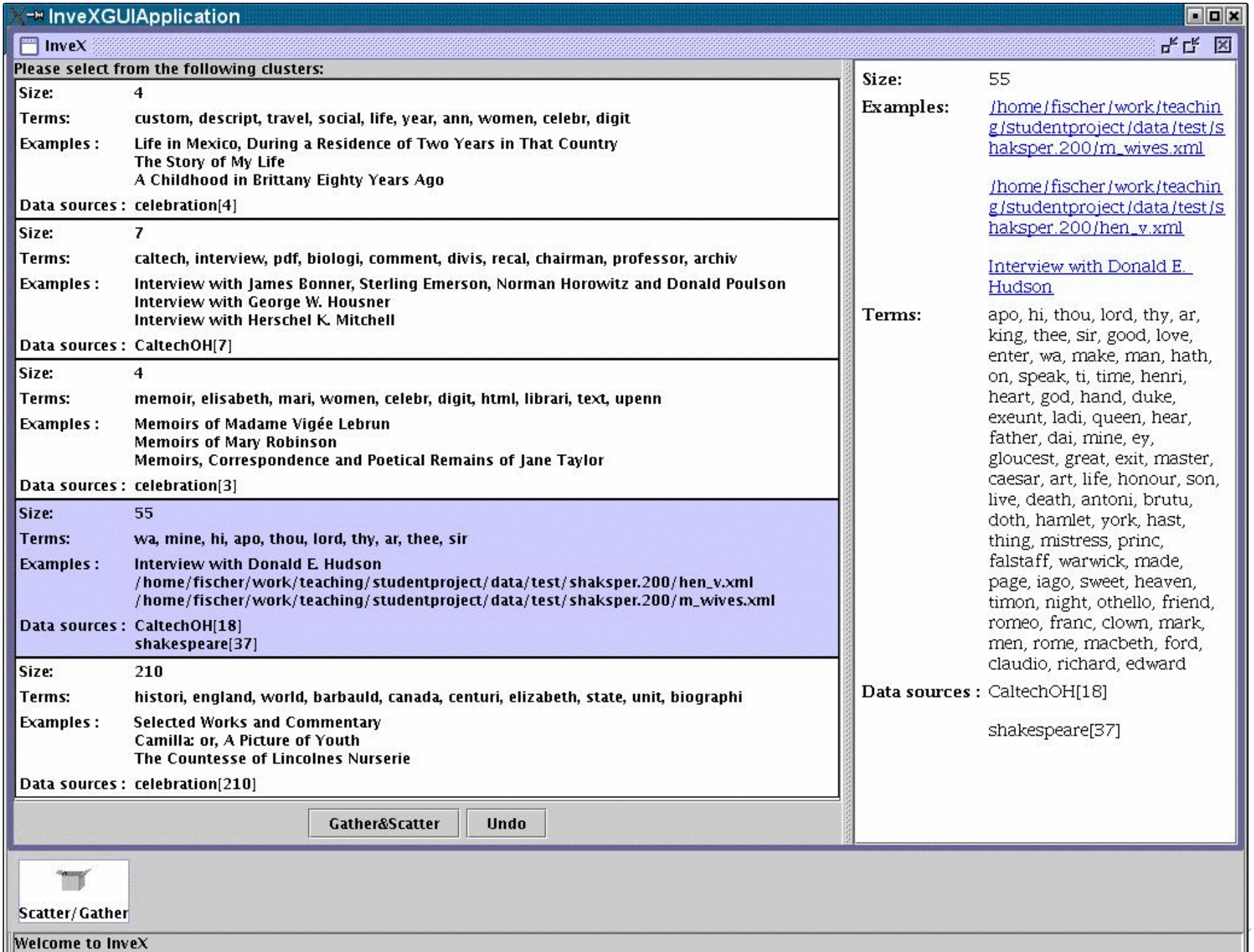
- 631459 "The Safra Affair: A Saga of Corporate Intrigue --- The Vendetta: How American Express Orchestrated a Smear Of Rival E
- 662803 "Kidder Advised U.S. It Was Helping BCCI Buy an Interest in First American ---- By Peter Truell Staff Reporter of The W
- 21620 "High Court Refuses to Dismiss Helmsley Indictment"
- 649610 "The Americas: Peru: Another Link in the BCCI Money Laundering Chain? ---- By Alvaro Vargas Llosa"
- 572658 "Senior Banker Charged In Money Laundering Operation"

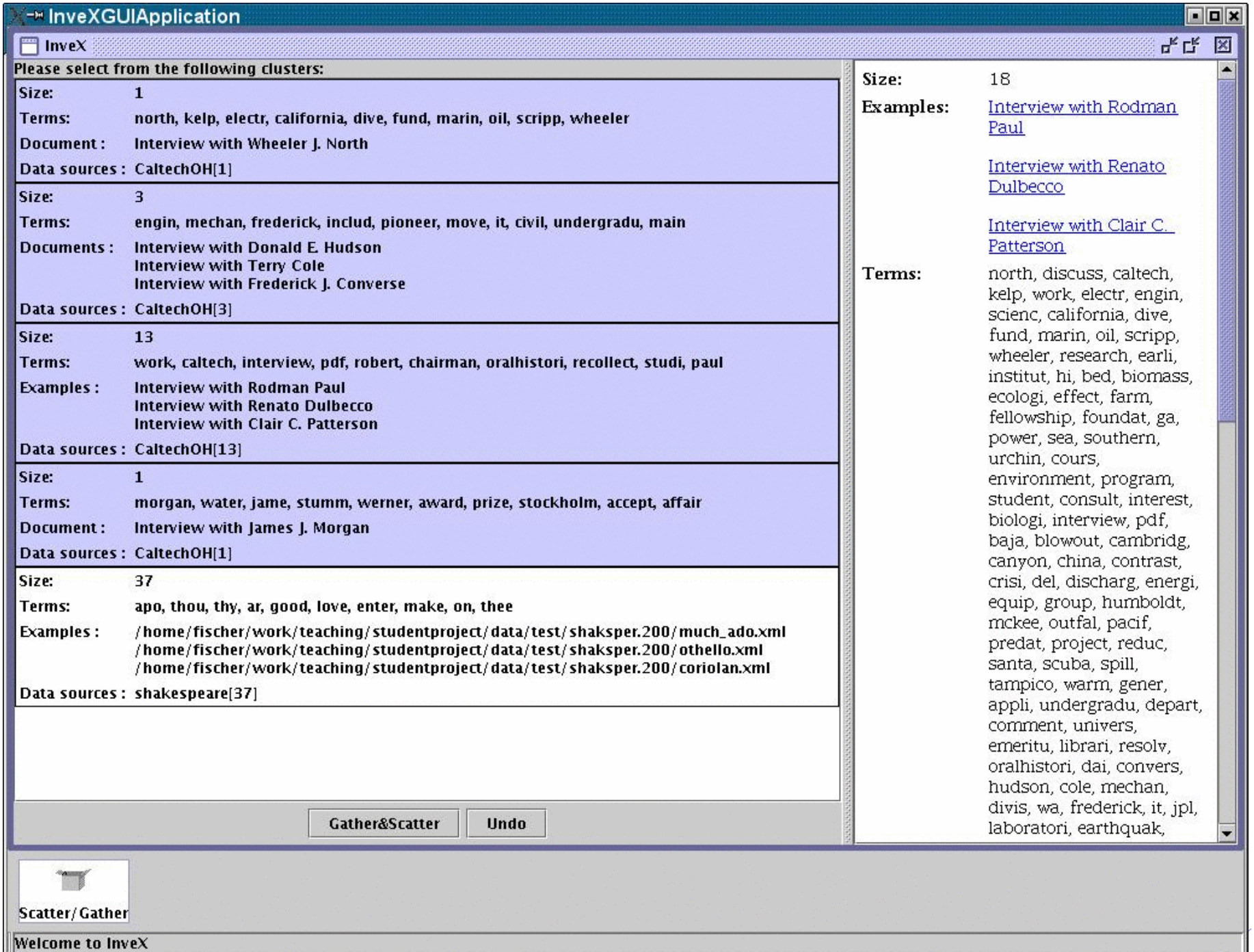
Cluster 5 Size: 7 marcos philippine marcoses unite order export respondent racketeering khashoggi buy manl

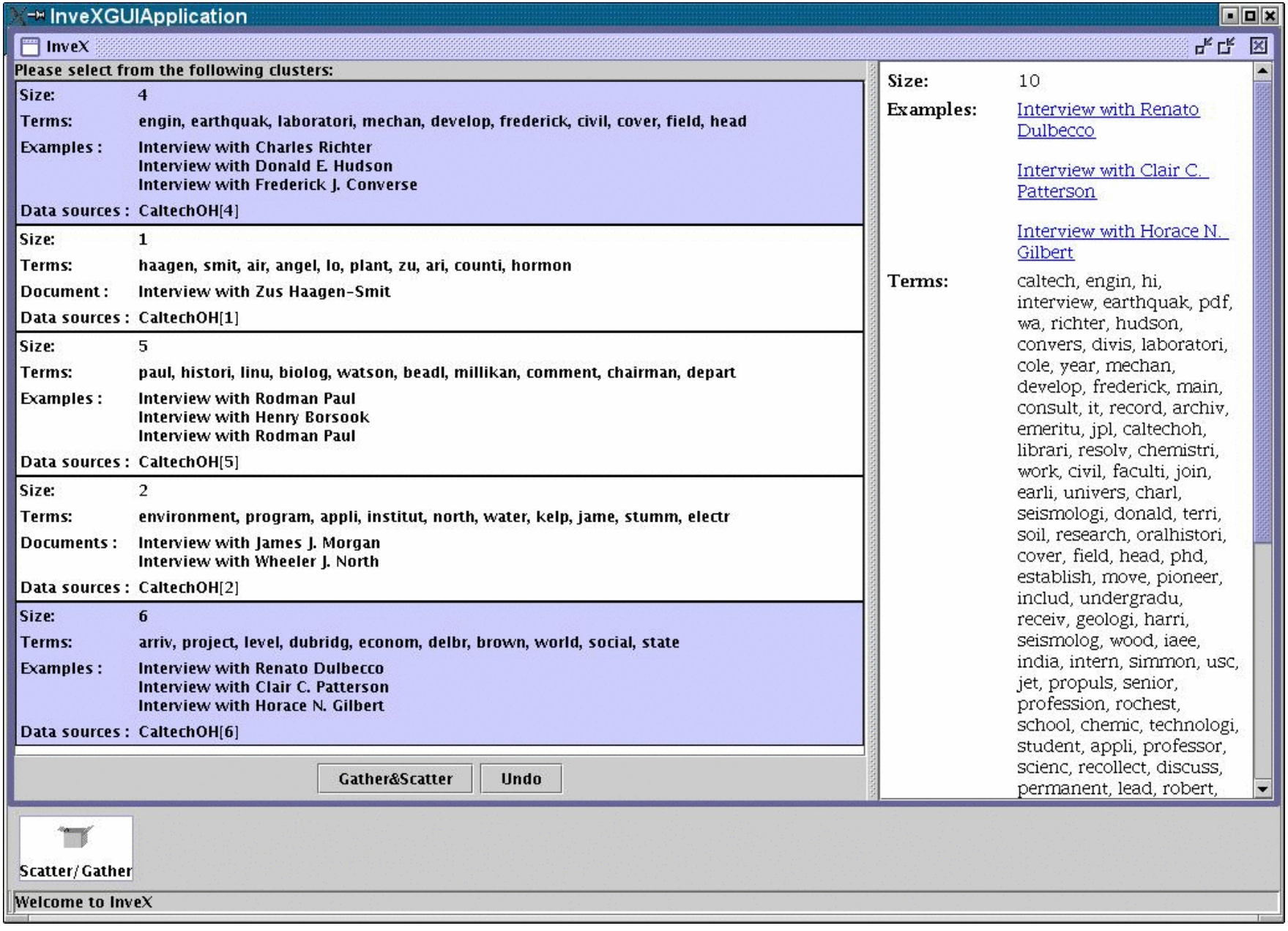
- 80628 "Former Interior Minister Extradited to Miami on Drug Charges"
- 37937 "Prosecutors Seek Judgment Against Marcos Even in Event of Death"
- 328041 "Action Affecting Export Privileges; Marek Cieslak"
- 575028 "Federal Grand Jury Indicts Marcos"

Studienprojekt Invisible Web (WS 03/04)
Scatter/Gather-Clustering für XML-Dokumente
aus dem Invisible Web
Beispiel...
Daten:

- *celebration – Werke englischer Autorinnen (Metadaten, aus Open Archives)*
- CaltechOH – Interviews mit Lehrenden an einer kalifornischen Universität (Metadaten, aus Open Archives)
- *shakespeare – Theaterstücke von Shakespeare in XML (Volltexte)*


InveXGUIApplication
InveX
Please select from the following clusters:
Size: 4
Terms: custom, descript, travel, social, life, year, ann, women, celebr, digit
Examples : Life in Mexico, During a Residence of Two Years in That Country
The Story of My Life
A Childhood in Brittany Eighty Years Ago
Data sources : celebration[4]
Size: 7
Terms: caltech, interview, pdf, biologi, comment, divis, recal, chairman, professor, archiv
Examples : Interview with James Bonner, Sterling Emerson, Norman Horowitz and Donald Poulson
Interview with George W. Housner
Interview with Herschel K. Mitchell
Data sources : CaltechOH[7]
Size: 4
Terms: memoir, elisabeth, mari, women, celebr, digit, html, librari, text, upenn
Examples : Memoirs of Madame Vigée Lebrun
Memoirs of Mary Robinson
Memoirs, Correspondence and Poetical Remains of Jane Taylor
Data sources : celebration[3]
Size: 55
Terms: wa, mine, hi, apo, thou, lord, thy, ar, thee, sir
Examples : Interview with Donald E. Hudson
/home/fischer/work/teaching/studentproject/data/test/shaksper.200/hen_v.xml
/home/fischer/work/teaching/studentproject/data/test/shaksper.200/m_wives.xml
Data sources : CaltechOH[18]
shakespeare[37]
Size: 210
Terms: histori, england, world, barbauld, canada, centuri, elizabeth, state, unit, biographi
Examples : Selected Works and Commentary
Camilla: or, A Picture of Youth
The Countesse of Lincolnes Nurserie
Data sources : celebration[210]
Gather&Scatter
Undo
Size: 55
Examples: /home/fischer/work/teaching/studentproject/data/test/shaksper.200/m_wives.xml
/home/fischer/work/teaching/studentproject/data/test/shaksper.200/hen_v.xml
Interview with Donald E. Hudson
Terms: apo, hi, thou, lord, thy, ar, king, thee, sir, good, love, enter, wa, make, man, hath, on, speak, ti, time, henri, heart, god, hand, duke, exeunt, ladi, queen, hear, father, dai, mine, ey, gloucest, great, exit, master, caesar, art, life, honour, son, live, death, antoni, brutu, doth, hamlet, york, hast, thing, mistress, princ, falstaff, warwick, made, page, iago, sweet, heaven, timon, night, othello, friend, romeo, franc, clown, mark, men, rome, macbeth, ford, claudio, richard, edward
Data sources : CaltechOH[18]
shakespeare[37]
Scatter/Gather
Welcome to InveX


InveXGUIApplication
InveX
Please select from the following clusters:
Size: 1
Terms: north, kelp, electr, california, dive, fund, marin, oil, scripp, wheeler
Document : Interview with Wheeler J. North
Data sources : CaltechOH[1]
Size: 3
Terms: engin, mechan, frederick, includ, pioneer, move, it, civil, undergradu, main
Documents : Interview with Donald E. Hudson
Interview with Terry Cole
Interview with Frederick J. Converse
Data sources : CaltechOH[3]
Size: 13
Terms: work, caltech, interview, pdf, robert, chairman, oralhistori, recollect, studi, paul
Examples : Interview with Rodman Paul
Interview with Renato Dulbecco
Interview with Clair C. Patterson
Data sources : CaltechOH[13]
Size: 1
Terms: morgan, water, jame, stumm, werner, award, prize, stockholm, accept, affair
Document : Interview with James J. Morgan
Data sources : CaltechOH[1]
Size: 37
Terms: apo, thou, thy, ar, good, love, enter, make, on, thee
Examples : /home/fischer/work/teaching/studentproject/data/test/shaksper.200/much_ado.xml
/home/fischer/work/teaching/studentproject/data/test/shaksper.200/othello.xml
/home/fischer/work/teaching/studentproject/data/test/shaksper.200/coriolan.xml
Data sources : shakespeare[37]
Gather&Scatter
Undo
Size: 18
Examples: Interview with Rodman Paul
Interview with Renato Dulbecco
Interview with Clair C. Patterson
Terms: north, discuss, caltech, kelp, work, electr, engin, scienc, california, dive, fund, marin, oil, scripp, wheeler, research, earli, institut, hi, bed, biomass, ecologi, effect, farm, fellowship, foundat, ga, power, sea, southern, urchin, cours, environment, program, student, consult, interest, biologi, interview, pdf, baja, blowout, cambridg, canyon, china, contrast, crisi, del, discharg, energi, equip, group, humboldt, mckee, outfal, pacif, predat, project, reduc, santa, scuba, spill, tampico, warm, gener, appli, undergradu, depart, comment, univers, emeritu, librari, resolv, oralhistori, dai, convers, hudson, cole, mechan, divis, wa, frederick, it, jpl, laboratori, earthquak,
Scatter/Gather
Welcome to InveX


InveXGUIApplication
InveX
Please select from the following clusters:
Size: 4
Terms: engin, earthquak, laboratori, mechan, develop, frederick, civil, cover, field, head
Examples : Interview with Charles Richter
Interview with Donald E. Hudson
Interview with Frederick J. Converse
Data sources : CaltechOH[4]
Size: 1
Terms: haagen, smit, air, angel, lo, plant, zu, ari, counti, hormon
Document : Interview with Zus Haagen-Smit
Data sources : CaltechOH[1]
Size: 5
Terms: paul, histori, linu, biolog, watson, beadl, millikan, comment, chairman, depart
Examples : Interview with Rodman Paul
Interview with Henry Borsook
Interview with Rodman Paul
Data sources : CaltechOH[5]
Size: 2
Terms: environment, program, appli, institut, north, water, kelp, jame, stumm, electr
Documents : Interview with James J. Morgan
Interview with Wheeler J. North
Data sources : CaltechOH[2]
Size: 6
Terms: arriv, project, level, dubridg, econom, delbr, brown, world, social, state
Examples : Interview with Renato Dulbecco
Interview with Clair C. Patterson
Interview with Horace N. Gilbert
Data sources : CaltechOH[6]
Gather&Scatter
Undo
Size: 10
Examples: Interview with Renato Dulbecco
Interview with Clair C. Patterson
Interview with Horace N. Gilbert
Terms: caltech, engin, hi, interview, earthquak, pdf, wa, richter, hudson, convers, divis, laboratori, cole, year, mechan, develop, frederick, main, consult, it, record, archiv, emeritu, jpl, caltechoh, librari, resolv, chemistri, work, civil, faculti, join, earli, univers, charl, seismologi, donald, terri, soil, research, oralhistori, cover, field, head, phd, establish, move, pioneer, includ, undergradu, receiv, geologi, harri, seismolog, wood, iaee, india, intern, simmon, usc, jet, propuls, senior, profession, rochest, school, chemic, technologi, student, appli, professor, scienc, recollect, discuss, permanent, lead, robert,
Scatter/Gather
Welcome to InveX

Clustering-Algorithmus: Buckshot
basiert auf K-Means
Ähnlichkeitsmaß: Cosinus zwischen Termgewichtvektoren
Cluster-Repräsentation:

- Titel von Dokumenten in der Nähe des Zentroiden
- wichtigste (gemäß der Termgewichtung) Terme im Cluster
- Datenherkunft

Gewicht eines Terms in einem Cluster C

tf_m	Häufigkeit des Terms im Dokument d_m
f_C	# Dokumente im Cluster C, in denen der Term vorkommt
f	# Dokumente insgesamt, in denen der Term vorkommt
n_C	# Dokumente im Cluster C
n	# Dokumente insgesamt
k	# Cluster in der Kollektion

Termgewichtungen:

- nach Häufigkeit des Terms im Cluster C:

$$\sum_{d_m \in C} tf_m$$

- nach relativem Informationsgehalt im Cluster:

$$f_C \cdot \left(\log \frac{f_C + \frac{1}{k}}{n_C + 1} - \log \frac{f - f_C + \frac{1}{k}}{n - n_C + 1} \right)$$