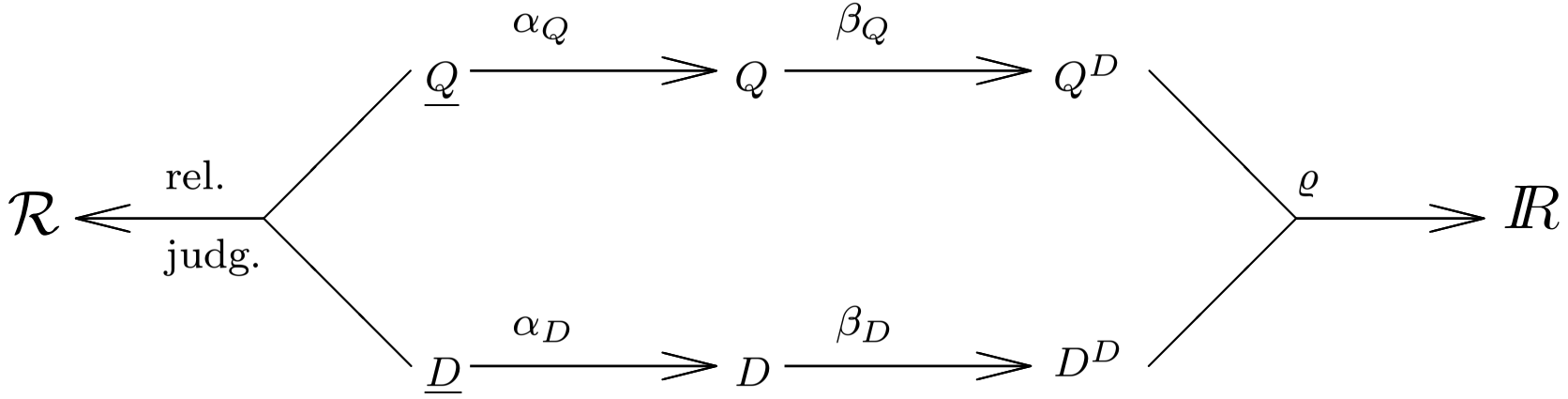
5 Nicht-Probabilistische Retrievalmodelle

- Boolesches Retrieval
- Fuzzy-Retrieval
- Vektorraummodell
- Clustering

5.1 Notationen



$\underline{Q}$: Informationsbedürfnisse

Q: Frage-Repräsentationen

Q^D: Frage-Beschreibungen (formale Anfragen)

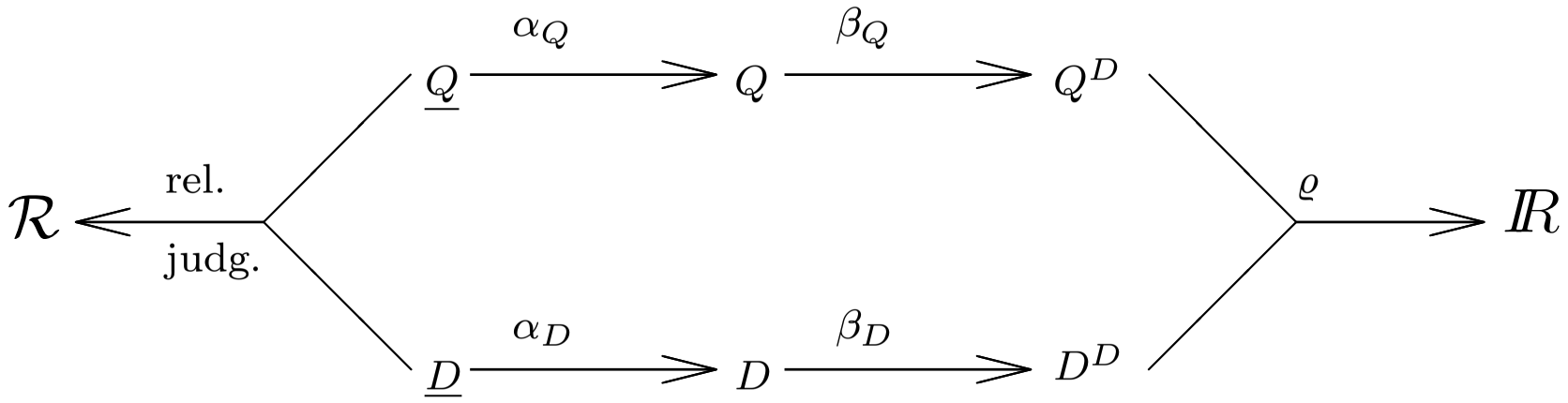
$\underline{D}$: Dokumente

D: Dokument-Repräsentationen

D^D: Dokument-Beschreibungen (Dok.-Indexierungen)

$\mathcal{R}$: Relevanzskala

ϱ: Retrievalfunktion



$T = \{t_1, \ldots, t_n\}$: Indexierungsvokabular

q_k: Frage-Repräsentation

q_k^D: Frage-Beschreibung

d_m: Dokument-Repräsentation

d_m^D: Dokument-Beschreibung $\vec{d}_m = (d_{m_1}, \ldots, d_{m_n})$: Dokument-Beschreibung als Menge von Indexierungsgewichten

$\mathcal{R}$: Relevanzskala

ϱ: Retrievalfunktion

5.2 Überblick über die Modelle

	Bool.	Fuzzy	Vektor	Prob.	Cluster.
theoretische Basis:					
– boolesche Logik	x				
– Fuzzy-Logik		x			
– Vektoralgebra			x		x
– Wahrsch.-Theorie				x	
Bezug zur Retrievalqual.		(x)		x	
gewichtete Indexierung		x	x	x	x
gewichtete Frageterms		(x)	x	x	
Fragestruktur:					
– linear			x	x	
– boolesch	x	x	(x)	(x)	
Suchmodus:					
– Suchen	x	x	x	x	
– Browsen					x

5.3 Boolesches Retrieval

Historisch als erstes Retrievalmodell entwickelt und eingesetzt
(Dokument-Beschreibungen auf Magnetbändern!)

Dokumenten-Beschreibungen D^D:
ungewichtete Indexierung, d.h. $d_m^D = \vec{d}_m$ mit $d_{m_i} \epsilon \{0, 1\}$ für $i = 1, \ldots, n$

boolesches Retrieval liefert nur Zweiteilung der Dokumente in „gefundene“ $(\varrho = 1)$
und „nicht gefundene“ $(\varrho = 0)$ Dokumente

Frage-Beschreibungen Q^D:

1. $t_i \, \epsilon T \Rightarrow t_i \, \epsilon Q^D$
2. $q_1, q_2 \, \epsilon \, Q^D \Rightarrow q_1 \wedge q_2 \, \epsilon \, Q^D$
3. $q_1, q_2 \, \epsilon \, Q^D \Rightarrow q_1 \vee q_2 \, \epsilon \, Q^D$
4. $q \epsilon Q^D \Rightarrow \neg q \, \epsilon \, Q^D$

Retrievalfunktion $\varrho(q, d_m)$:

1. $t_1 \, \epsilon T \Rightarrow \varrho(t_i, \vec{d}_m) = d_{m_i}$
2. $\varrho(q_1 \wedge q_2, \vec{d}_m) = \min(\varrho(q_1, \vec{d}_m), \varrho(q_2, \vec{d}_m))$
3. $\varrho(q_1 \vee q_2, \vec{d}_m) = \max(\varrho(q_1, \vec{d}_m), \varrho(q_2, \vec{d}_m))$
4. $\varrho(\neg q, \vec{d}_m) = 1 - \varrho(q, \vec{d}_m)$

Mächtigkeit der booleschen Anfragesprache:

jede beliebige Dokumentenmenge kann selektiert werden

Voraussetzung: alle Dokumente besitzen unterschiedliche Indexierungen
Konstruktion der booleschen Frageformulierung q_k zu einer vorgegebenen Dokumentenmenge D_k:

$$
\begin{aligned}
d_m^Q &= x_{m_1} \wedge \ldots \wedge x_{m_n} \text{ mit} \\
x_{m_i} &= \begin{cases} t_i & \text{falls } d_{m_i} = 1 \\ \neg t_i & \text{sonst} \end{cases} \\
q_k &= \bigvee_{d_j \epsilon D_k} d_j^Q
\end{aligned}
$$

Beispiel-Recherche

"The side effects of drugs on memory or cognitive abilities, not related to aging"

1.	19248	DRUGS
2.	2412	DRUGS in TI
3.	2560	AGING
4.	19119	DRUG not AGING
5.	2349	#2 and #4
6.	9305	MEMORY
7.	6	#5 and (DRUG near4 MEMORY)
8.	22091	COGNITIVE
9.	16	#5 and (DRUG near4 COGNITIVE)
10.	22	#7 or #9
11.	2023	SIDE-EFFECTS-DRUG in DE
12.	0	#11 and #10

Nachteile des booleschen Retrieval

1. Größe der Antwortmenge ist schwierig zu kontrollieren
2. Keine Ordung der Antwortmenge nach mehr oder weniger relevanten Dokumenten
3. Keine Möglichkeit zur Gewichtung von Fragetermen oder gewichteter Indexierung
4. Trennung in gefundene und nicht gefundene Dokumente zu streng:
   Zu $q = t_1 \wedge t_2 \wedge t_3$ werden Dokumente mit zwei gefundenen Termen genauso zurückgewiesen wie solche mit 0
   Analog für $q = t_1 \vee t_2 \vee t_3$ keine Unterteilung der gefundenen Dokumente
5. Erstellung der Frageformulierung sehr umständlich
6. schlechte Retrievalqualität

5.4 Fuzzy-Retrieval

Teilweise Überwindung der Nachteile des booleschen Retrieval

Dokumenten-Beschreibungen:
Erweiterung auf gewichtete Indexierung, d.h. $d_{m_i} \epsilon [0,1]$

Frage-Beschreibungen, Retrievalfunktion:
wie beim booleschen Retrieval

Retrievalfunktion liefert jetzt Werte $\varrho(q_k^D, \vec{d}_m) \epsilon [0,1]$
$\rightarrow$ Ranking der Antwortmenge

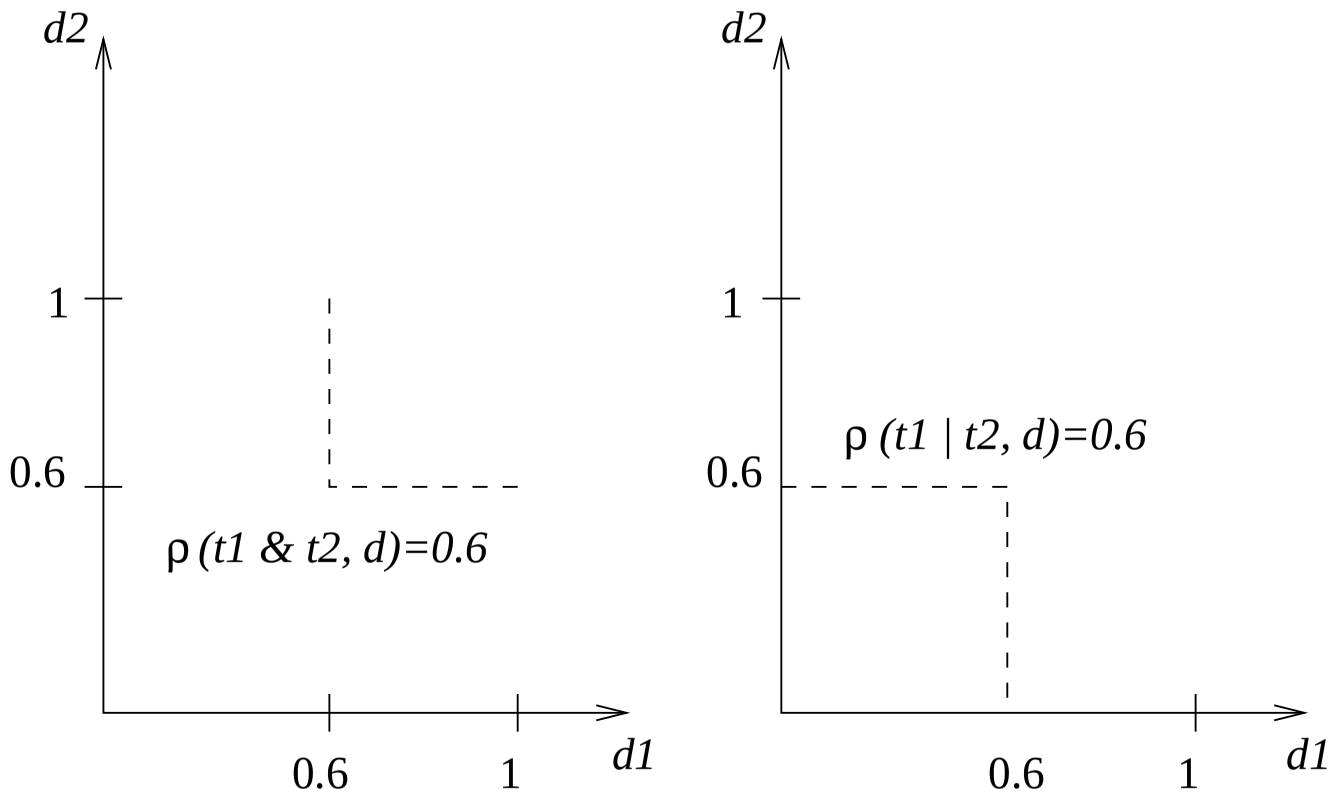
aber: Definition der Retrievalfunktion umstritten

$$T = \{t_1, t_2\}$$

$$q = t_1 \wedge t_2$$

$$\vec{d}_1 = (0.6, 0.6) \quad , \quad \vec{d}_2 = (0.59, 0.99)$$

$$\varrho(q, \vec{d}_1) = 0.6 \quad , \quad \varrho(q, \vec{d}_2) = 0.59$$



Andere Definitionen der Fuzzy-Operatoren (T-Normen)
überwinden Nachteile der Standard-Definition,
aber verletzen Gesetze der Booleschen Algebra:
(z.B. $\varrho(((t_1 \vee t_2) \wedge t_3), d) \neq \varrho(((t_1 \wedge t_3) \vee (t_2 \wedge t_3)), d))$

Kollektion	MEDLARS	ISI	INSPEC	CACM
#Dok.	1033	1460	12684	3204
#Fragen	30	35	77	52
Bool.	0.2065		0.1159	
Fuzzy	0.2368	0.1000	0.1314	0.1551
Vektor	0.5473	0.1569	0.2325	0.3027

Experimenteller Vergleich von Booleschem Retrieval, Fuzzy-Retrieval und Vektorraummodel

Beurteilung des Fuzzy-Retrieval

+ Generalisierung des booleschen Retrieval für gewichtete Indexierung → Ranking
– keine Fragetermgewichtung
– schlechte Retrievalqualität
– Erstellung der Frageformulierung sehr umständlich

5.5 Das Vektorraummodell

zuerst entstanden im Rahmen der Arbeiten zu SMART
(experimentelles Retrievalsystem von G. Salton und Mitarbeitern (Harvard/Cornell), seit 1961)
in den 80er Jahren von Wong und Raghavan überarbeitet

Dokumente und Fragen als Punkte in einem orthonormalen Vektorraum, der durch die Terme aufgespannt wird

orthonormaler Vektorraum:

- alle Term-Vektoren orthogonal (und damit auch linear unabhängig)
- alle Term-Vektoren normiert

Dokument-Beschreibung: ähnlich wie Fuzzy-Retrieval
$d_m^D = \vec{d}_m$ mit $d_{m_i} \epsilon I\!R$ für $i = 1, \ldots, n$

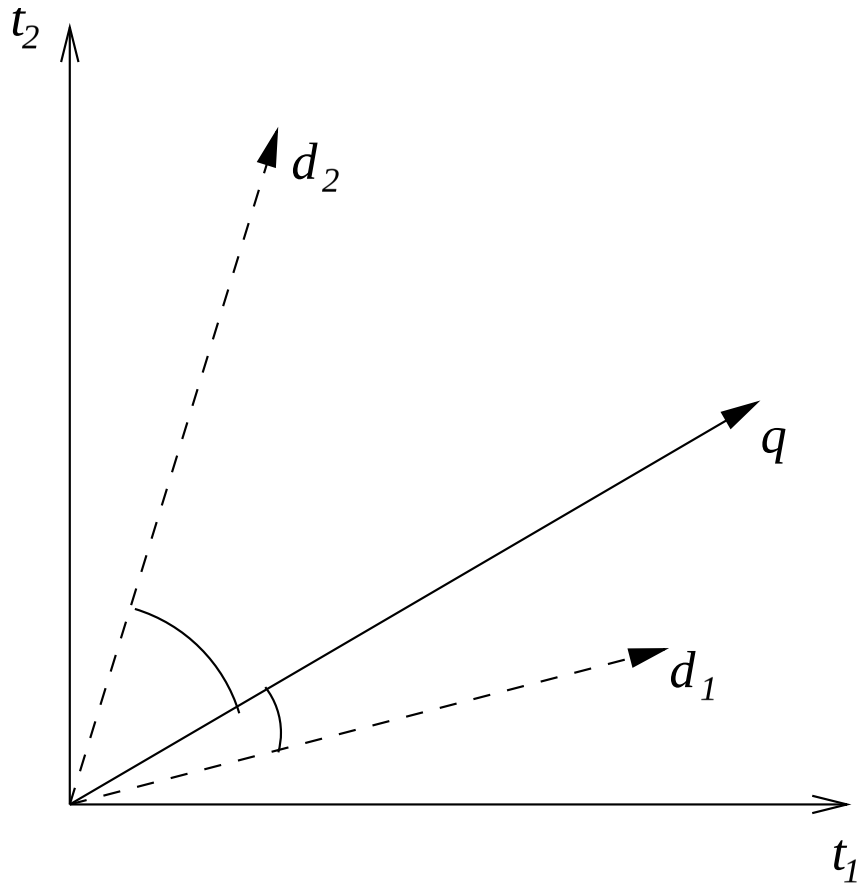
Frage-Beschreibung:
$q_k^Q = \vec{q}_k$ mit $q_{k_i} \epsilon I\!R$ für $i = 1, \ldots, n$

Vektorraummodell: Retrievalfunktion

Vektor-Ähnlichkeitsmaße, z.B. Cosinus

Meistens: Skalarprodukt

$\varrho(\vec{q}_k, \vec{d}_m) = \vec{q}_k \cdot \vec{d}_m$



Beispiel-Frage: „side effects of drugs on memory and cognitive abilities“

t_i	q_{k_i}	d_{1_i}	d_{2_i}	d_{3_i}	d_{4_i}
side effect	2	1	0.5	1	1
drugs	2	1	1	1	1
memory	1	1		1	
cognitive ability	1		1	1	0.5
Retrievalgewicht		5	2.5	6	4.5

Coordination Level Match

Vereinfachung des Vektorraummodells:
nur binäre Frage- und Dokumenttermgewichtung

Dokument-Beschreibung: ähnlich wie Boolesches Retrieval
$d_m^D = \vec{d}_m$ mit $d_{m_i} \varepsilon \{0, 1\}$ für $i = 1, \ldots, n$

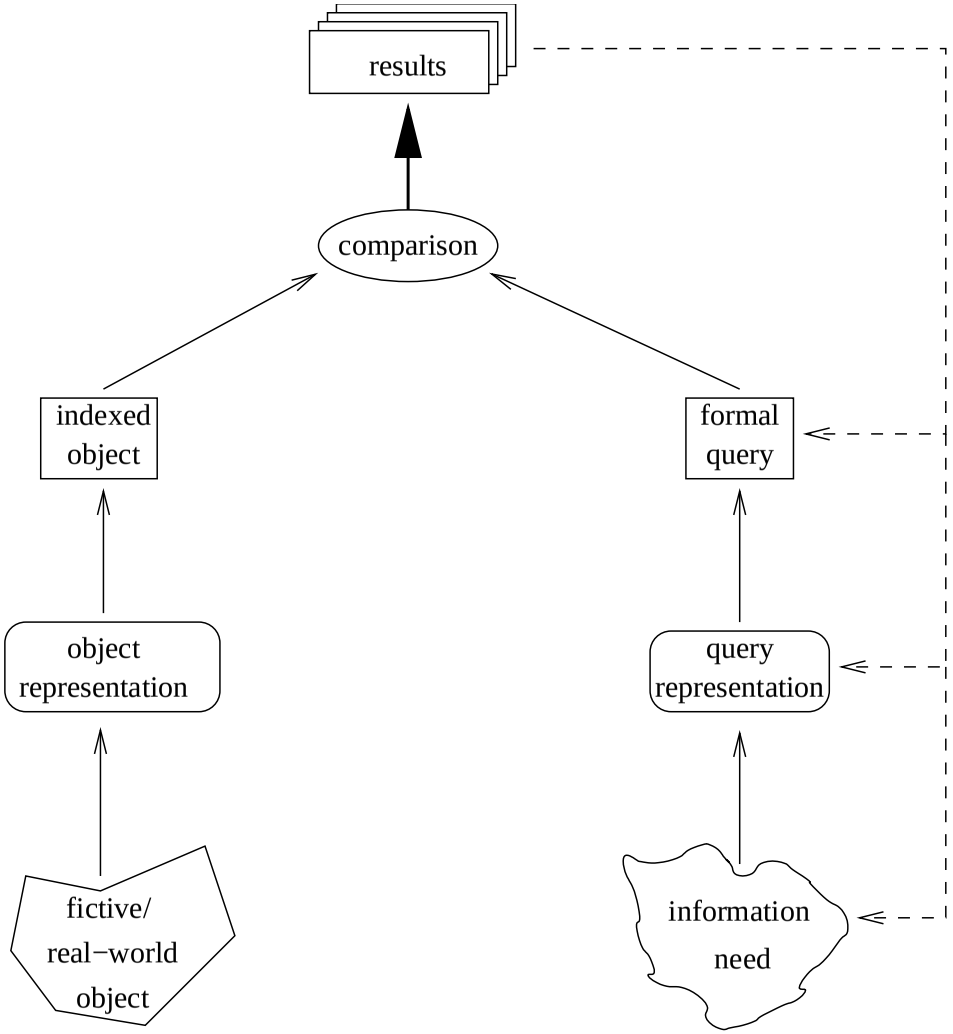
Frage-Beschreibung:
$q_k^Q = \vec{q}_k$ mit $q_{k_i} \varepsilon \{0, 1\}$ für $i = 1, \ldots, n$

Retrievalfunktion:
Skalarprodukt

$$\varrho(\vec{q}_k, \vec{d}_m) = \vec{q}_k \cdot \vec{d}_m = |q_k^T \cap d_m^T|$$

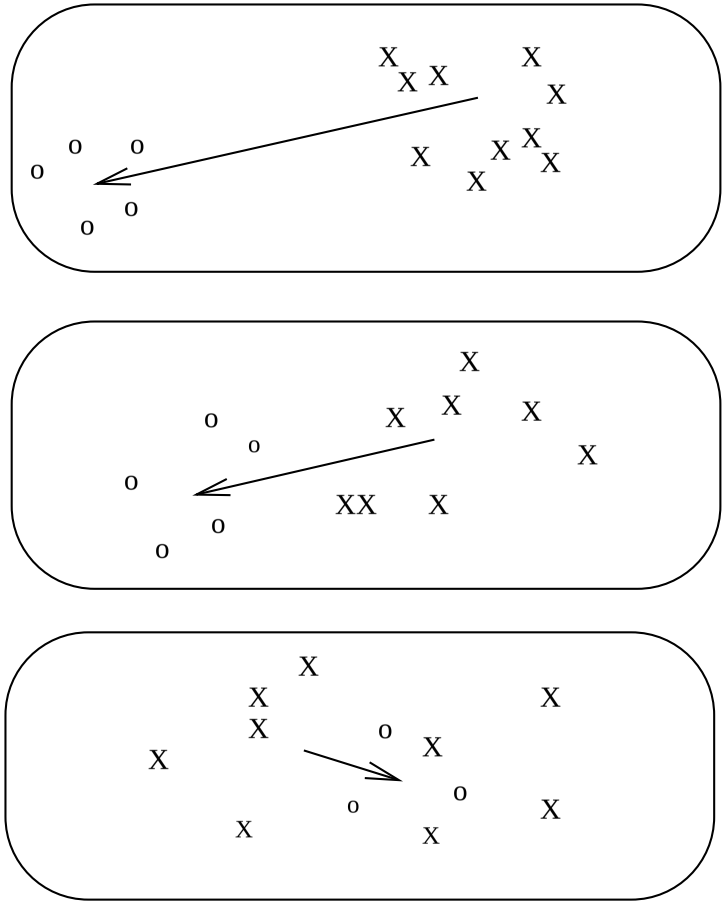
5.5.1 Relevance Feedback

iteratives Retrieval:



Relevance Feedback im VRM

Ziel: Modifikation des Fragevektors



D^R: relevante Dokumente

D^N: irrelevante Dokumente

Idee:
wähle Fragevektor $\vec{q}$ so, dass Differenz der RSVs zwischen relevanten und irrelevanten Dokumenten maximal wird:

$$\sum_{(d_k,d_l)\in D^R\times D^N} \vec{q}\vec{d_k} - \vec{q}\vec{d_l} \stackrel{!}{=} \max$$

mit der Nebenbedingung

$$\sum_{i=1}^{n} q_i^2 = c$$

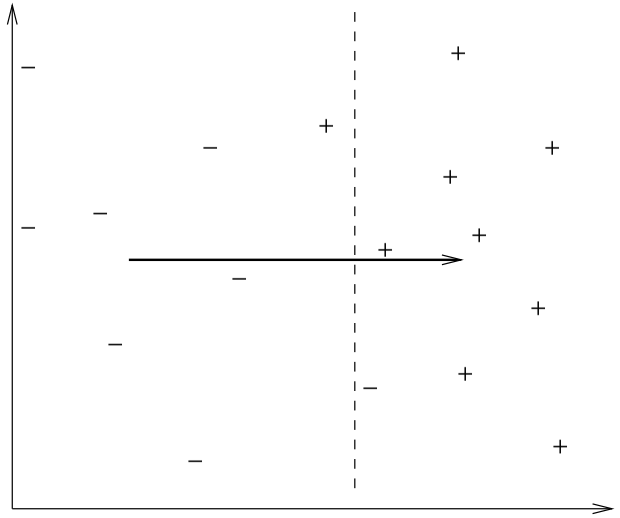
Extremwertproblem mit Randbedingung
$\rightarrow$ Lagrange-Multiplikator einsetzen

$$
\begin{aligned}
F &= \lambda\left(\sum_{i=1}^{n} q_i^2 - c\right) + \sum_{(d_k,d_l)\in D^R\times D^N}\sum_{i=1}^{n} q_i d_{k_i} - q_i d_{l_i} \\
\frac{\partial F}{\partial q_i} &= 2\lambda q_i + \sum_{(d_k,d_l)\in D^R\times D^N} d_{k_i} - d_{l_i} \stackrel{!}{=} 0 \\
q_i &= -\frac{1}{2\lambda}\sum_{(d_k,d_l)\in D^R\times D^N} d_{k_i} - d_{l_i} \\
\vec{q} &= -\frac{1}{2\lambda}\sum_{(d_k,d_l)\in D^R\times D^N} \vec{d_k} - \vec{d_l} \\
&= -\frac{1}{2\lambda}|D^N|\sum_{d_k\in D^R}\vec{d_k} - |D^R|\sum_{d_l\in D^N}\vec{d_l} \\
&= -\frac{|D^N||D^R|}{2\lambda}\frac{1}{|D^R|}\sum_{d_k\in D^R}\vec{d_k} - \frac{1}{|D^N|}\sum_{d_l\in D^N}\vec{d_l}
\end{aligned}
$$

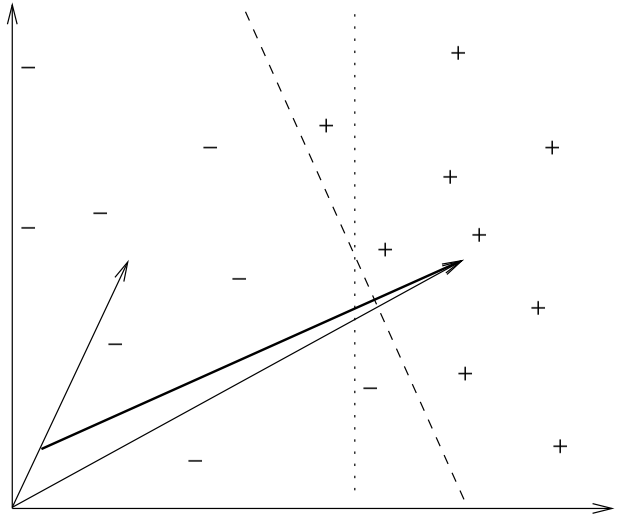
wähle c so, dass $|D^N||D^R|/2\lambda = -1$:

$$\vec{q} = \frac{1}{|D^R|} \sum_{d_k \in D^R} \vec{d_k} - \frac{1}{|D^N|} \sum_{d_l \in D^N} \vec{d_l}$$

$\hat{=}$ Verbindungsvektor der Zentroiden der relevanten / irrelevanten Dokumente



unterschiedliche Gewichtung positiver und negativer Beispiele:



Rocchio-Algorithmus

- unterschiedliche Gewichtung positiver und negativer Beispiele
- Berücksichtigung der ursprünglichen Anfrage

$$\vec{q_k}' = \vec{q_k} + \alpha \frac{1}{|D_k^R|} \sum_{d_j \epsilon D_k^R} \vec{d_j} - \beta \frac{1}{|D_k^N|} \sum_{d_j \epsilon D_k^N} \vec{d_j}$$

α, β — positive Konstanten, heuristisch festzulegen (z.B. $\alpha = 0.75$, $\beta = 0.25$)

Vorgehensweise:

1. Retrieval mit Fragevektor $\vec{q_k}$ vom Benutzer
2. Relevanzbeurteilung der obersten Dokumente der Rangordnung
3. Berechnung eines verbesserten Fragevektors $\vec{q_k}'$ aufgrund der Feedback-Daten
4. Retrieval mit dem verbesserten Vektor
5. Evtl. Wiederholung der Schritte 2-4

5.5.2 Dokumenten-Indexierung

Vektorraum-Modell liefert keine Aussagen darüber, wie die Dokumenten-Indexierung zu berechnen ist!

(Dokumenten-)Indexierung im Vektoraummodell:
heuristische Formeln zur Berechnung der Indexierungsgewichte

zugrundeliegende Dokumenten-Repräsentation: Multi-Menge (Bag) von Terms

d_m^T	Menge der in d_m vorkommenden Terms
l_m	Dokumentlänge (# laufende Wörter in d_m)
al	durchschnittliche Dokumentlänge in $\underline{D}$
tf_{mi}:	Vorkommenshäufigkeit (Vkh) von t_i in d_m.
n_i:	# Dokumente, in denen t_i vorkommt.
$\|\underline{D}\|$:	# Dokumente in der Kollektion

inverse Dokumenthäufigkeit (idf):

$$idf_i = \frac{\log \frac{|\underline{D}|}{n_i}}{|\underline{D}| + 1}$$

normalisierte Vorkommenshäufigkeit:

$$ntf_i = \frac{tf_{mi}}{tf_{mi} + 0.5 + 1.5\frac{l_m}{al}}$$

Indexierungsgewicht tfidf:

$$w_{mi} = ntf_i \cdot idf_i$$

Kollektion	CACM	CISI	CRAN	INSPEC	MED
Coord.	0.185	0.103	0.241	0.094	0.413
SMART	0.363	0.219	0.384	0.263	0.562

Binäre Gewichte (Coordination Level Match) vs. SMART-Gewichtung von Fragen und Dokumenten
(aus Salton/Buckley 88)

Beurteilung des Vektorraummodells

+ einfaches Modell, insbes. für den Benutzer
+ unmittelbar anwendbar auf neue Kollektionen
+ gute Retrievalqualität
– sehr viele heuristische Komponenten
– kein Bezug zur Retrievalqualität
  (Optimalität von Relevance Feedback?)
– Dokumentrepräsentation kann schlecht erweitert werden

5.6 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

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:
   $$\min_{d_i \in C_l} sim(d_k, d_i) \leq \alpha$$
   b) complete link-Clustering:
   $$\max_{d_i \in C_l} sim(d_k, d_i) \leq \alpha$$
   c) average link-Clustering:
   $$\frac{1}{|C_l|} \sum_{d_i \in C_l} sim(d_k, d_i) \leq \alpha$$
3. falls es kein solches Cluster gibt, bildet d_k ein neues Cluster.

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

partitionierendes Clustering

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

Aufwand: $O(kn)$

aber: Ergebnis hängt stark von der Wahl der seed-Dokumente ab!

hierarchisches Clustering

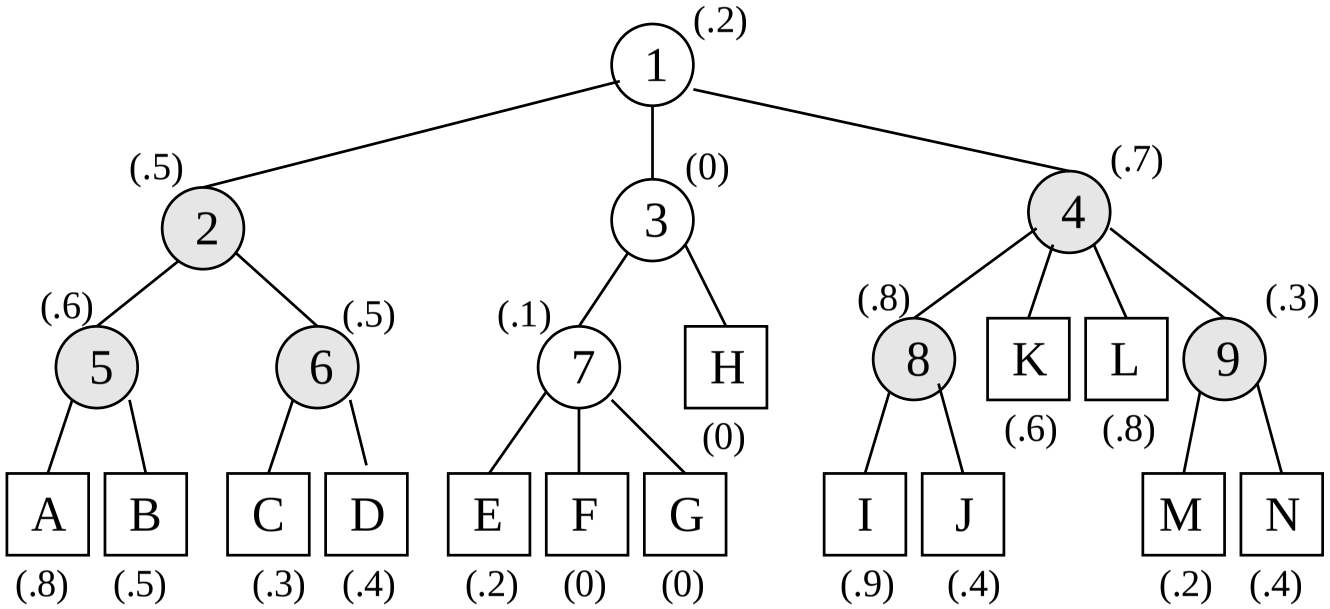
liefert baumförmige Cluster-Struktur

Realisierung:
iterative Anwendung der o.g. Clustering-Verfahren
zur weiteren Zerlegung der gebildeten Cluster

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 von Dokumenten

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:

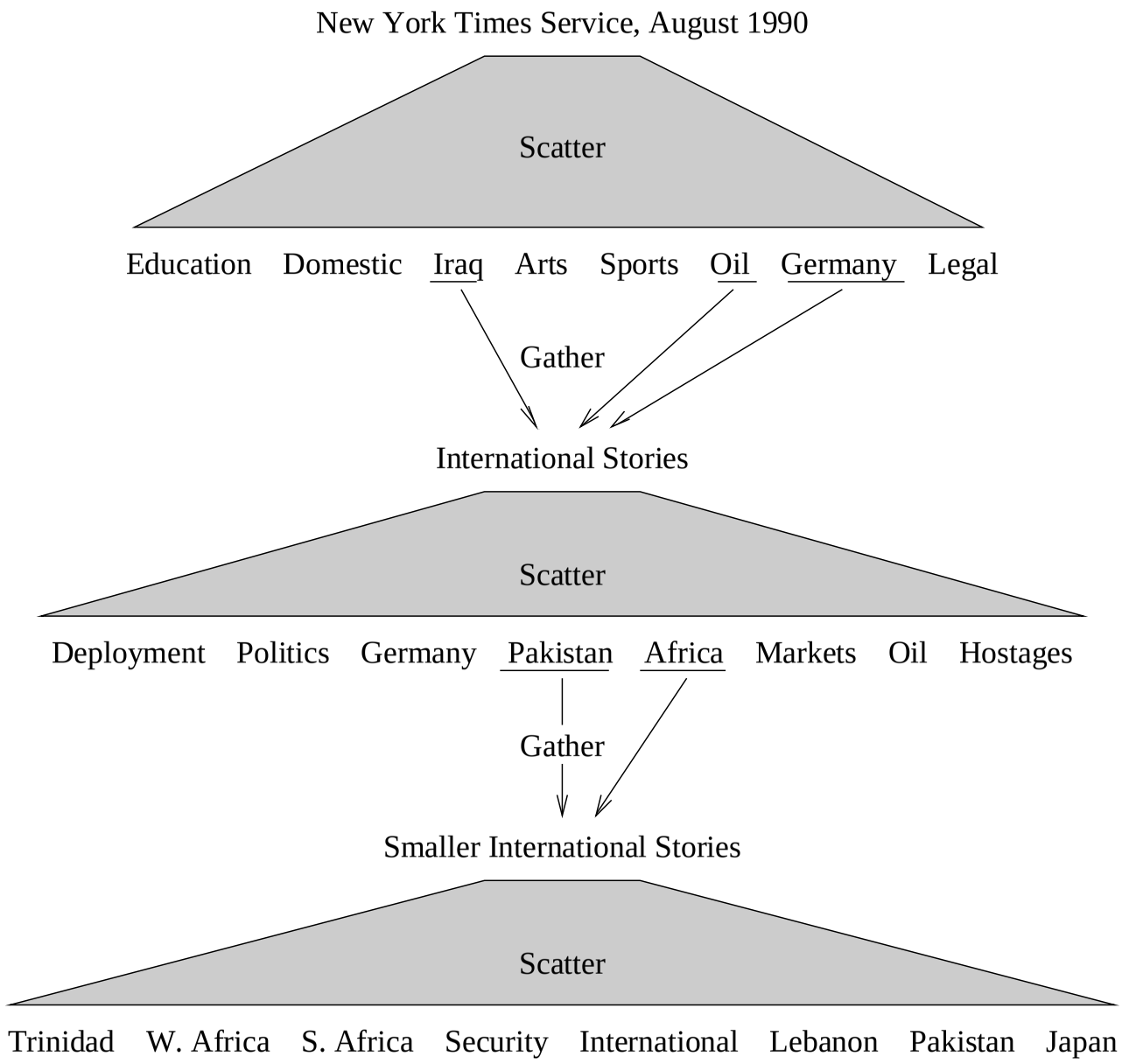
- Ähnlichkeitssuche sinnvoll als Ergänzung zu den anderen Retrievalmodellen (es werden andere relevante Dokumente gefunden)
- Clustering ermöglicht Browsing
- Vorprozessierung der Cluster nur für Retrieval lohnt nicht

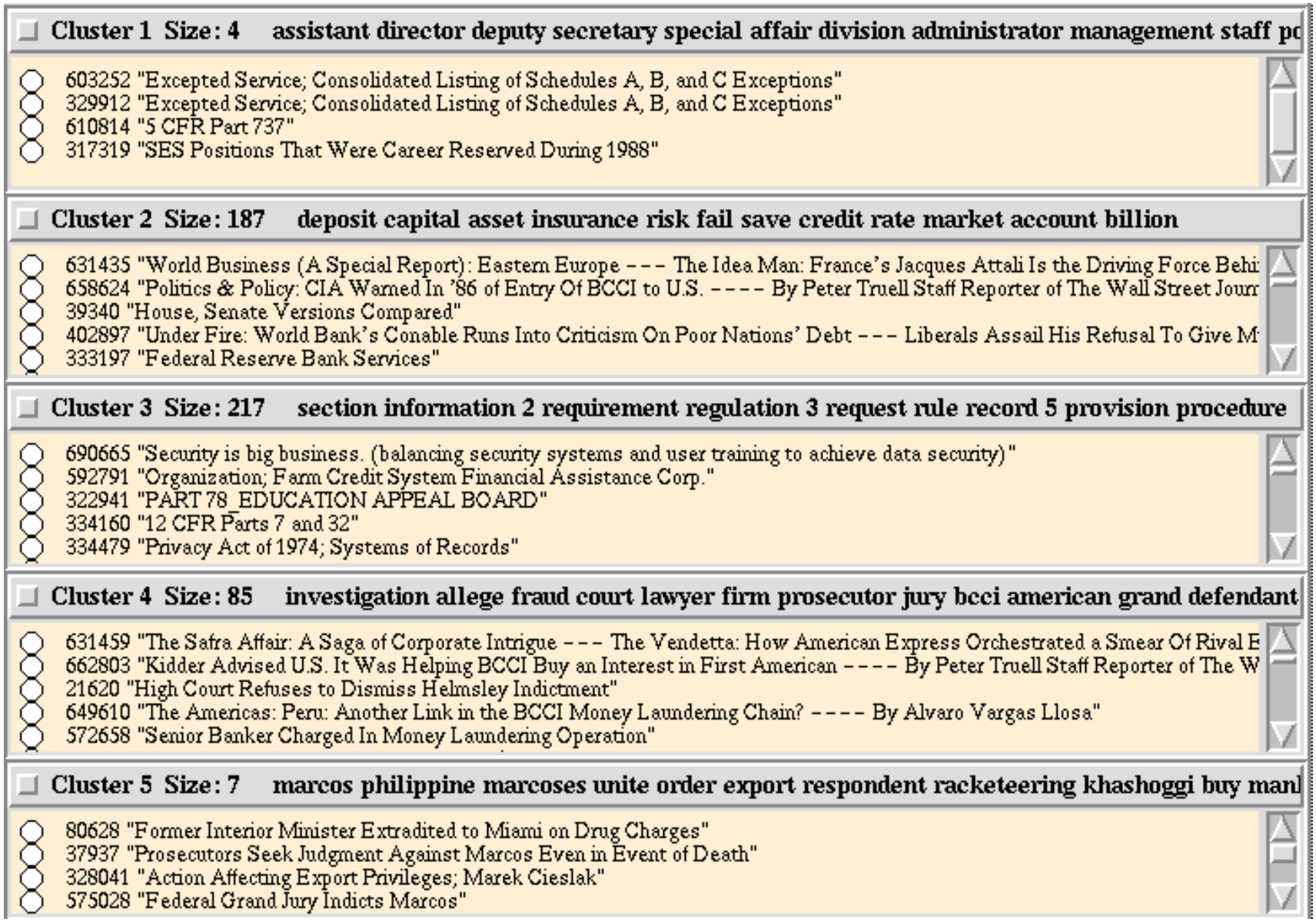
Scatter-Gather-Clustering

basiert auf partitionierendem Clustering

Cluster-Repräsentation:

- Titel von Dokumenten in der Nähe des Zentroiden
- häufige Wörter im Cluster


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 Jourr
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"