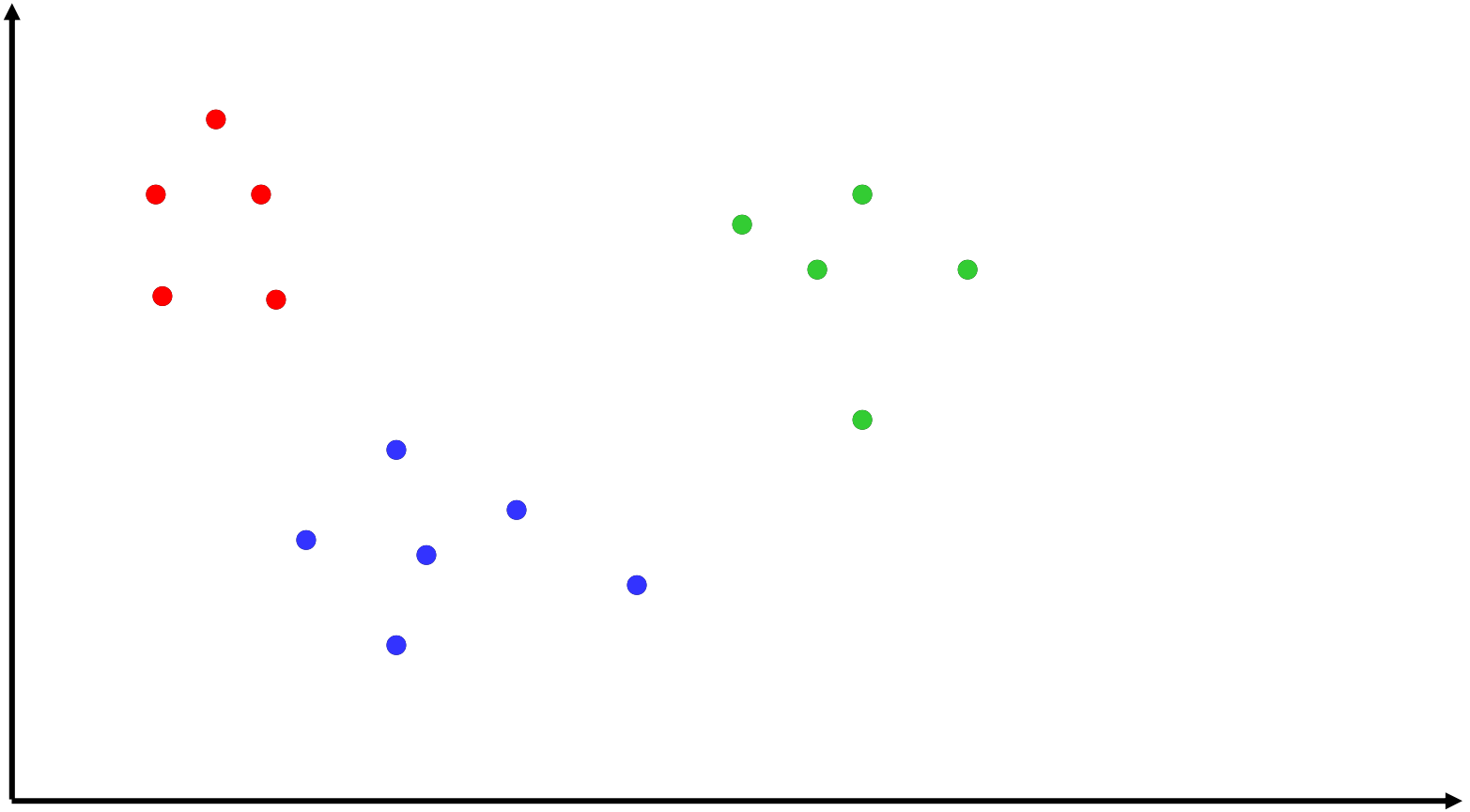
Text Clustern

Clustern

- Teile nicht kategorisierte Beispiele in disjunkte Untermengen, so genannte *Cluster,* ein, so daß:
  - Beispiele innerhalb eines Clusters sich sehr ähnlich
  - Beispiele in verschiedenen Clustern möglichst sehr unterschiedlich sind.
- Entdecke neue Kategorien in einer *unüberwachten* Art
- es werden keine Schlagwort für die Kategorien vorab zur Verfügung gestellt

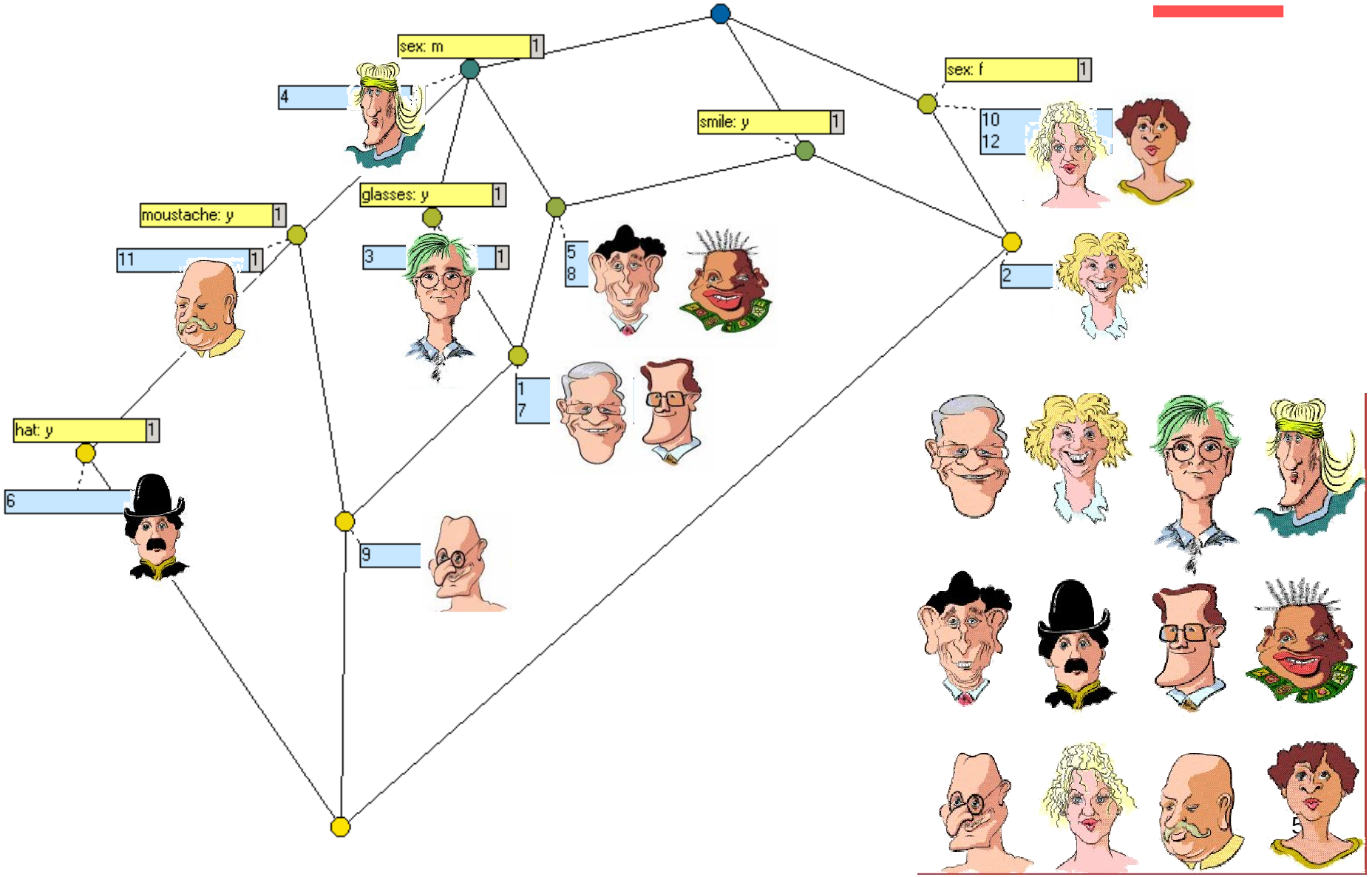
Einführung Clustern



Einführung Clustern

case	sex	glasses	moustache	smile	hat
1	m	y	n	y	n
2	f	n	n	y	n
3	m	y	n	n	n
4	m	n	n	n	n
5	m	n	n	y?	n
6	m	n	y	n	y
7	m	y	n	y	n
8	m	n	n	y	n
9	m	y	y	y	n
10	f	n	n	n	n
11	m	n	y	n	n
12	f	n	n	n	n

Einführung Clustern



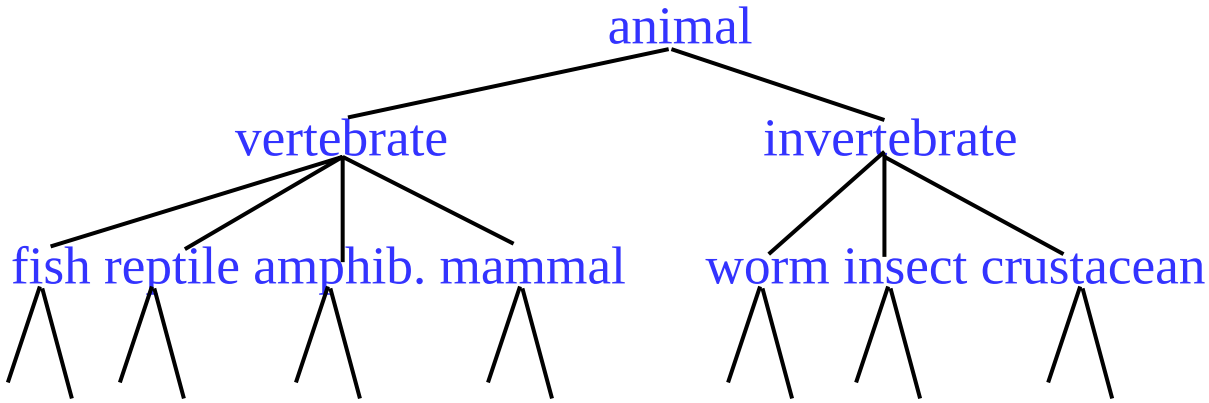
Typen von Clustering-Verfahren

- Hierarchische Verfahren
  - Parameter: Distanzfunktion für Punkte und für Cluster
  - bestimmt Hierarchie von Clustern, mischt jeweils die ähnlichsten Cluster
- Partitionierende Verfahren
  - Parameter: Anzahl *k* der Cluster, Distanzfunktion
  - sucht ein „flaches“ Clustering in *k* Cluster mit minimalen Kosten
- Dichtebasierte Verfahren
  - Parameter: minimale Dichte in einem Cluster, Distanzfunktion
  - erweitert Punkte um ihre Nachbarn solange Dichte groß genug
- Andere Clustering-Verfahren
  - Fuzzy Clustering
  - Soft Clustering (EM)
  - Graph-theoretische Verfahren
  - neuronale Netze

Mehr Details zu Clusterverfahren gibt es in der KDD Vorlesung.

Hierarchisches Clustern

- Bilde eine baum-basierte hierarchische Taxonomy (*Dendrogram*) aus einer Menge von Beispielen.



- Die rekursive Anwendung eines Standard Cluster Algorithmus kann hierarchisches Clustern bewirken (z.B. Bi-Sec-KMeans).

Aglommeratives vs. divisives Clustern

- *Aglommerative* (bottom-ip) Methoden beginnen mit je einem Beispiel als eigener Cluster und verbinden diese iterativ, um größere Cluster zu bilden.
- *Divisive* (*partitionierende*, *top-down*) trennen die Menge aller Beispiele in eine gegebene Anzahl von Cluster und wiederholen dies für jeden Cluster solange bis jeder Cluster nur noch ein Beispiel enthält.

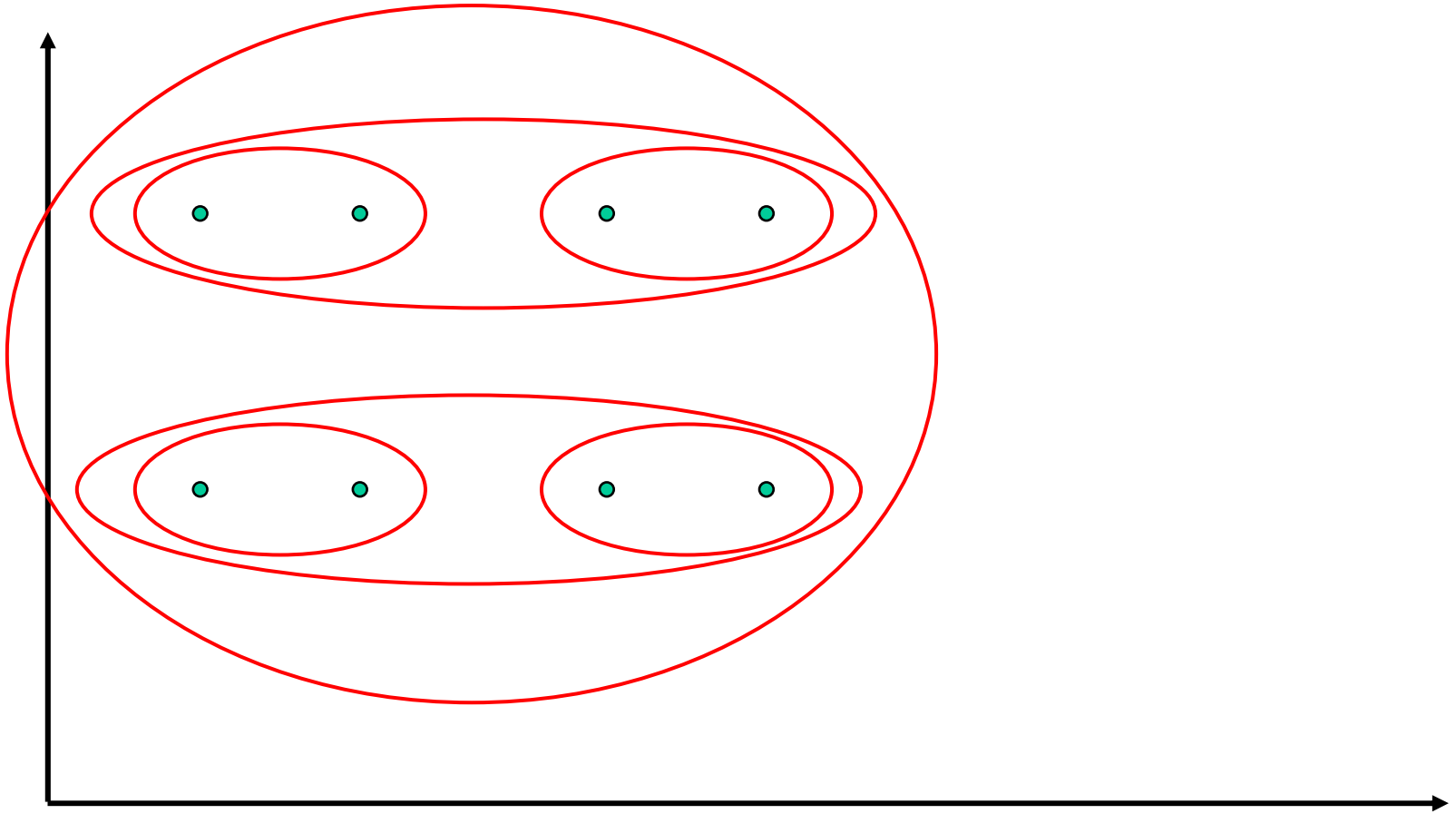
Hierarchisches Agglomeratives Clustern (HAC)

- Gegeben ist eine *Ähnlichkeitsfunktion* zur Bestimmung der Ähnlichkeit von zwei Objekten.
- Beginnt mit allen Objekten in einem separatem Cluster und verbindet dann mehrmals die beiden Cluster, die am ähnlichsten sind, bis nur noch ein Cluster da ist.
- Die Historie des Zusammenlegens bildet einen binären Baum oder eine Hierarchie.

Cluster Ähnlichkeit

- Ähnlichkeitsfunktion, die die Ähnlichkeit von zwei Objekten bestimmt : *sim*(*x*,*y*).
  - Kosinus Ähnlichkeit von Dokumentvektoren.
- Wie berechnet man die Ähnlichkeit von zwei Clustern, von denen jeder möglicherweise eine Vielzahl von Objekten enthält?
  - Single Link: Ähnlichkeit der zwei ähnlichsten Mitglieder.
  - Complete Link: Ähnlichkeit der zwei am wenigstens ähnlichen Mitglieder.
  - Average Link: Durchschnittsähnlichkeit zwischen allen Mitgliedern.

Single Link am Beispiel



Rechnerische Komplexität

- Bei der ersten Iteration, müssen alle HAC Methoden die Ähnlichkeit aller Paare von n Objekten berechnen. Aufwand: $O(n^2)$
- Vor jedem der nachfolgenden $n-2$ Zusammenlegungsschritte, muss der Abstand zwischen dem neu erzeugten Cluster und allen anderen noch existierenden Clustern berechnet werden.
- Um bei einem Gesamtaufwand von $O(n^2)$ zu bleiben, muss die Berechnung der Ähnlichkeit mit jedem Cluster in konstanter Zeit erfolgen.

Nicht-Hierarchisches Clustern

- Typischerweise muß man bei den meisten Verfahren die Anzahl der gewünschten Cluster, *k* angeben.
- Wähle willkürlich *k* Objekte als *Saat* (Ausgangspunkt) der Clusterung (einen pro Cluster).
- Bilde anfängliche Cluster, die auf dieser Saat basieren.
- Iteriere mehrfach und ordne Objekte Clustern neu zu, mit dem Ziel das Gesamtclusterergebnis zu verbessern.
- Stoppe, wenn das Clustern konvergiert oder nach einer festen Anzahl von Iterationen.

K-Means

- Gegeben: Objekte werden durch reelle-wertige Vektoren repräsentiert.
- Die Cluster basieren auf dem *Schwerpunkt* (Centroid) - dem Mittelwert von Punkten eines Cluster, *c*:

$$\vec{\mu}(c) = \frac{1}{|c|} \sum_{\vec{x} \in c} \vec{x}$$

- Die Neuzuordnung von Objekten zu Clustern basiert auf dem Abstand zu den aktuellen Cluster-Centroiden.

Abstandsmaße

- Euklidischer Abstand (L_2 Norm):

$$L_2(\vec{x},\vec{y}) = \sum_{i=1}^{m} (x_i - y_i)^2$$

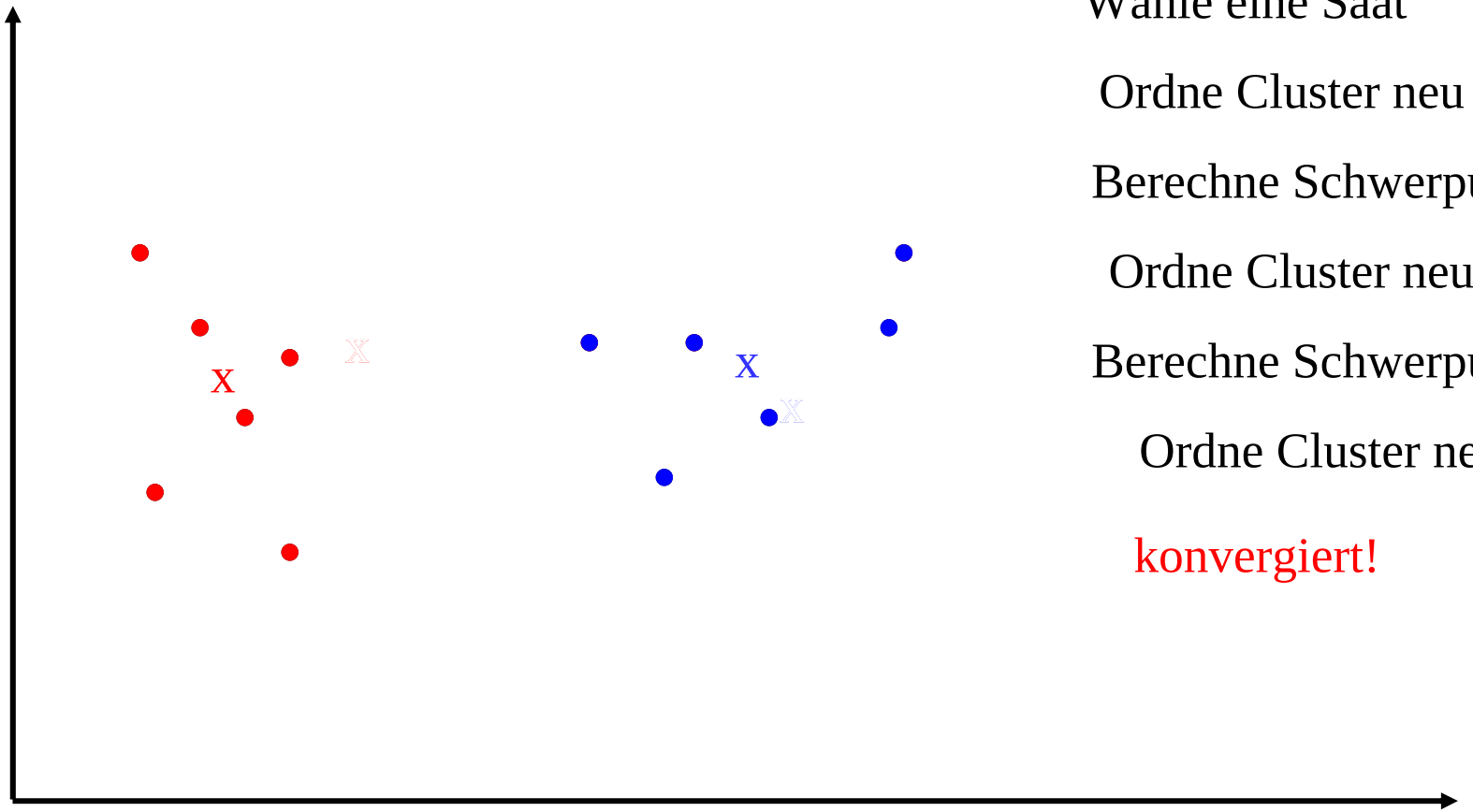
- L_1 Norm:

$$L_1(\vec{x},\vec{y}) = \sum_{i=1}^{m} |x_i - y_i|$$

- Kosinus Ähnlichkeit (→ Abstandsmaß durch Subtraktion von 1):

$$1 - \frac{\vec{x} \bullet \vec{y}}{|\vec{x}| \cdot |\vec{y}|}$$

K Means Beispiel (K=2)



Wähle eine Saat

Ordne Cluster neu zu

Berechne Schwerpunkte

Ordne Cluster neu zu

Berechne Schwerpunkte

Ordne Cluster neu zu

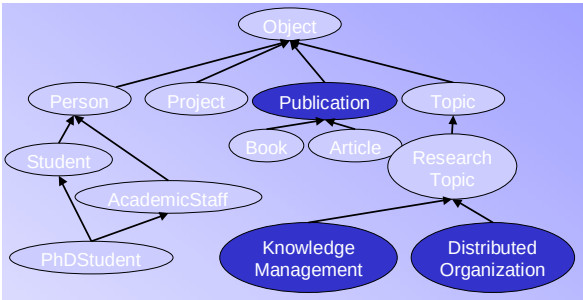
konvergiert!

Zeit Komplexität

- Der Aufwand zur Berechnung des Abstandes zwischen zwei Objekten sei O(*m*), wobei *m* die Dimensionalität der Vektoren ist.
- Neuzuordnung von Clustern: O(*kn*) Abstands-berechnungen, oder O(*knm*).
- Centroidberechnung: Jeder Objektvektor wird einmal zu seinem Schwerpunkt addiert: O(*nm*).
- Die letzten zwei Schritte werden jeweils einmal pro Iterationen *I* durchgeführt: O(*Iknm*).
- Linear in allen relevanten Faktoren, wobei von einer festen Anzahl von Iterationen ausgegangen wird,
- KMeans ist effizienter als HAC (O(n^2)).

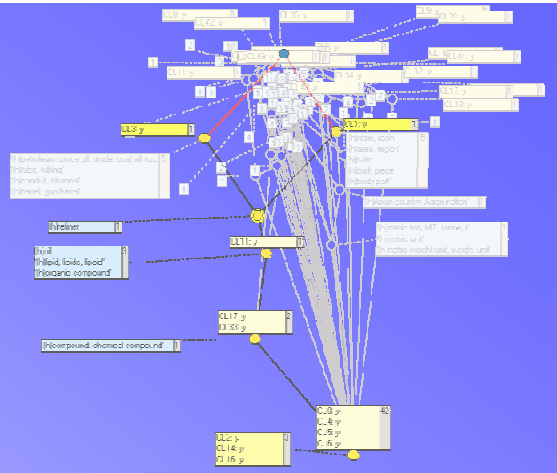
Text Clustering

- HAC und K-Means wurden zum Clustern von Texten wie folgt angewendet.
- Typischer Weise nutzt man einen ***normalisierten***, TF/IDF-gewichteten Vektor und die Kosinus-Ähnlichkeit.
- Die Berechung wird für dünn besetzte Vektoren optimiert.
- Anwendungen:
  - Bei einer typischen Suchanfrage werden Dokumente des gleichen Clusters passend zu der ursprünglichen Antwortmenge zurückgeliefert, um so den Recall zu erhöhen.
  - Clustern der Suchergebnisse um besser organisierte Ergebnisse dem Anwender anbieten zu können. (z.B. http://de.vivisimo.com/).
  - Die automatische Erstellung von taxonomischen Hierarchien für eine Menge von Dokumenten mit dem Ziel das Browsing zu erleichtern. (z.B. Yahoo & DMOZ).



CLUSTERCOUNT	60
EXAMPLE	100
MINCOUNT	15

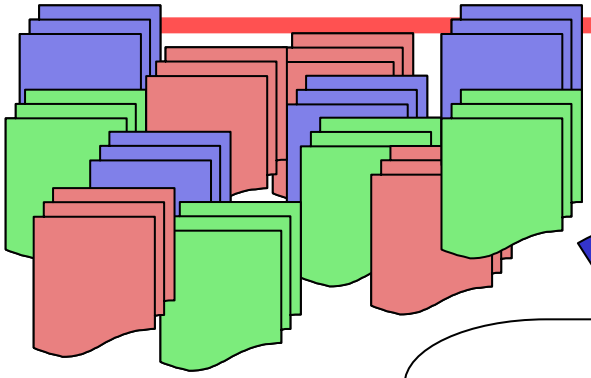
Mittelwert - PURITY				WEIGHT PRUNE					
				thilf			without		
ONTO	HYPDEP	HYPDI	HYPINT	30	5	0	30	5	0
false	0	context	add	0,570	0,547	0,526	0,470	0,461	0,462
true	0	context	repl	0,572	0,552	0,526	0,462	0,467	0,459
			add	0,585	0,560	0,546	0,483	0,484	0,483
			only	0,603	0,581	0,571	0,485	0,478	0,481
		first	repl	0,562	0,550	0,525	0,469	0,469	0,466
			add	0,590	0,563	0,551	0,486	0,482	0,486
			only	0,599	0,584	0,569	0,483	0,487	0,477
		all	repl	0,512	0,504	0,492	0,472	0,469	0,472
			add	0,533	0,530	0,525	0,481	0,480	0,484
			only	0,500	0,501	0,497	0,455	0,456	0,454
	5	context	repl	0,602	0,595	0,578	0,404	0,396	0,398
			add	0,618	0,598	0,586	0,423	0,413	0,407
			only	0,593	0,594	0,588	0,392	0,387	0,390
		first	repl	0,602	0,602	0,584	0,411	0,400	0,401
			add	0,616	0,600	0,586	0,423	0,417	0,411
			only	0,593	0,593	0,585	0,394	0,389	0,390
		all	repl	0,522	0,521	0,519	0,318	0,315	0,306
			add	0,542	0,537	0,531	0,324	0,318	0,309
			only	0,520	0,513	0,513	0,315	0,311	0,308



Reales Beispiel von Clustern von Texten mit Hintergrundwissen

Motivation

Objekte



Repräsentation der Objekte

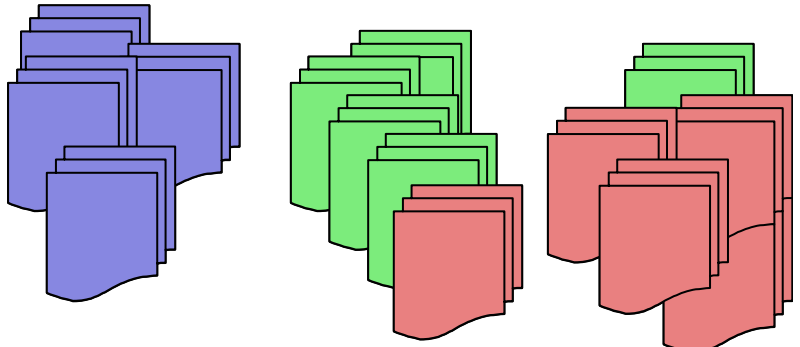
	Morgens	Abends	team	baseman
Obj1	1	1		
Obj2	1	1		
Obj3			2	1
Obj4			2	1

Hintergrund-wissen

nlichkeitsmaß
tanzfunktion

sterverfahren

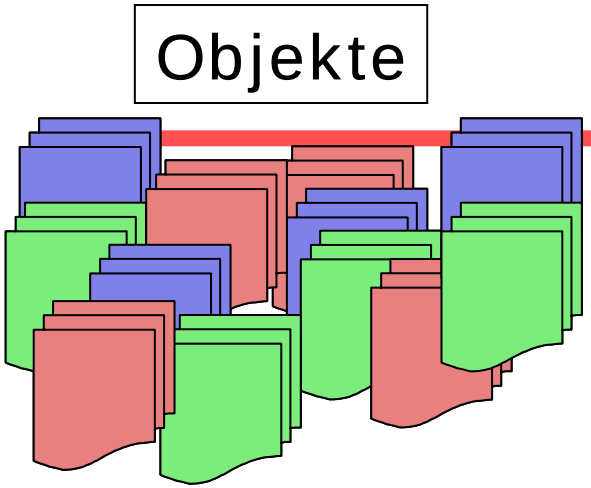
Erklärung



(Bi-Sec)-k-Means

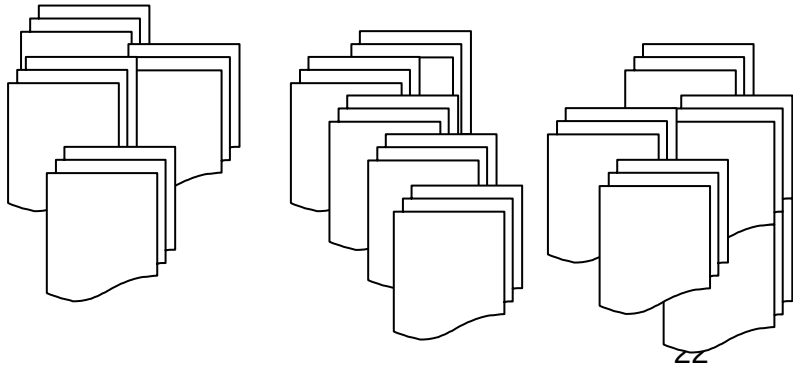
- Variante von k-Means
  - k_b Cluster werden berechnet
  - Wiederholte Anwendung des k-Means mit k=2 Clustern bis k_b Cluster berechnet sind
  - Größter Cluster wird geteilt
- Bessere Ergebnisse als k-Means im Bereich Text-Clustern
- Bessere Performanz als k-Means

Textdokumentclustern



Datensatz – Reuters

- Finanztexte des Jahres 1987
- 12344 Dokumente
- 82 Reuters-Themen/Klassen
- Teilmenge mit min. 15 Dokumenten pro Klasse und max. 100 Dokumenten



Vorverarbeitungsschritte

- Ableiten der Dokument-Term-Matrix ("Bag of Words")
  - Stoppworte entfernen
  - Wortstämme und deren Häufigkeiten *tf* berechnen
  - Gewichten der Dokument-Vektoren mit *tfidf* (term frequency / inverted document frequency)
  - 863167 Worte insgesamt
  - 20574 verschiedene Wortstämme im Datensatz

$$tfidf(d,t) = tf(d,t) \cdot \log\left(\frac{|D|}{df(t)}\right)$$

$|D|$ Anzahl Dokumente d
$df(t)$ Anzahl Dokumente d, in denen Term t vorkommt

Textdokumentclustern

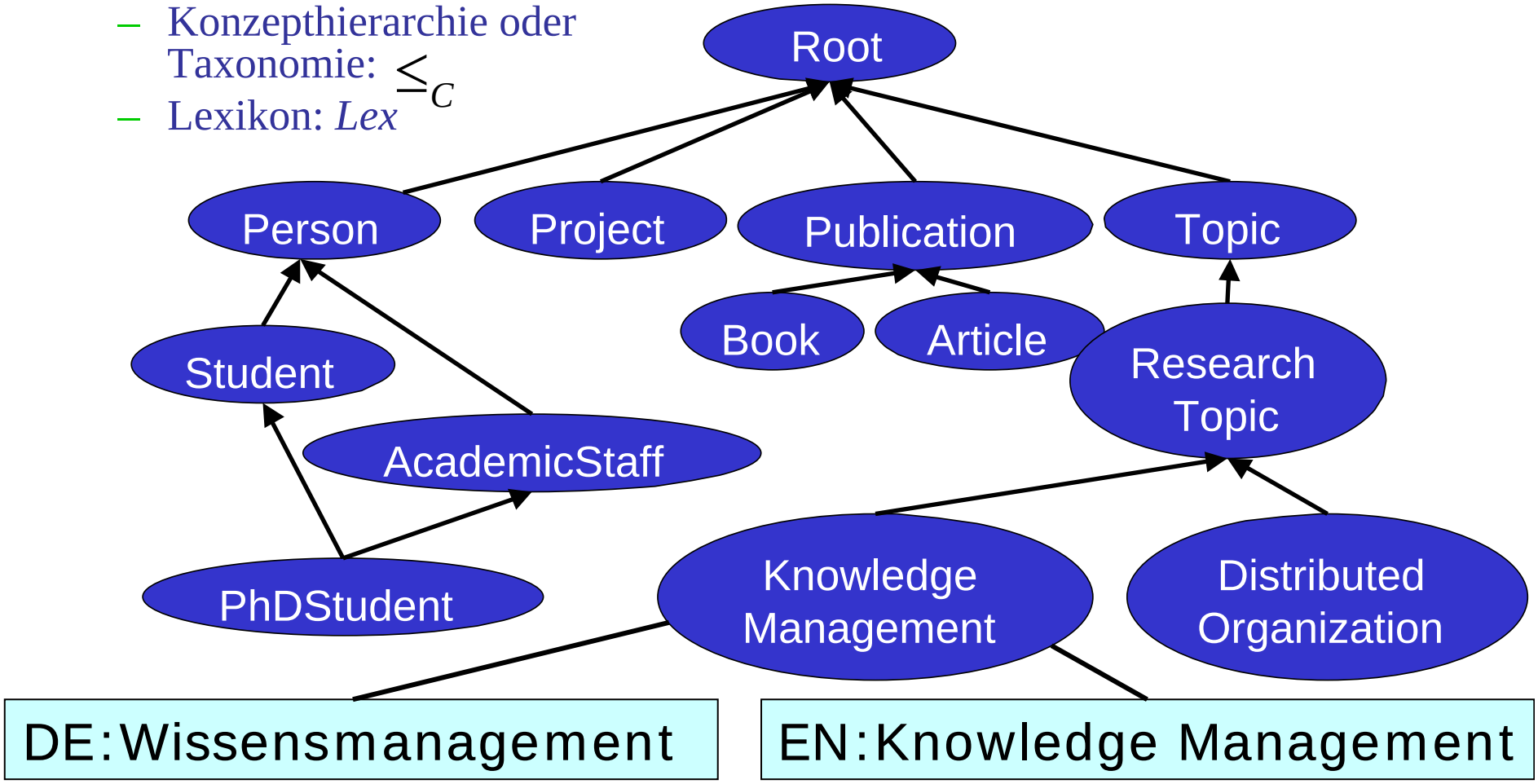
Dok 17892 crude
==========
Oman has granted term crude oil customers retroactive discounts from official prices of 30 to 38 cents per barrel on liftings made during February, March and April, the weekly newsletter Middle East Economic Survey (MEES) said. MEES said the price adjustments, arrived at through negotiations between the Omani oil ministry and companies concerned, are designed to compensate for the difference between market-related prices and the official price of 17.63 dlrs per barrel adopted by non-OPEC Oman since February.
REUTER

Bag of Words

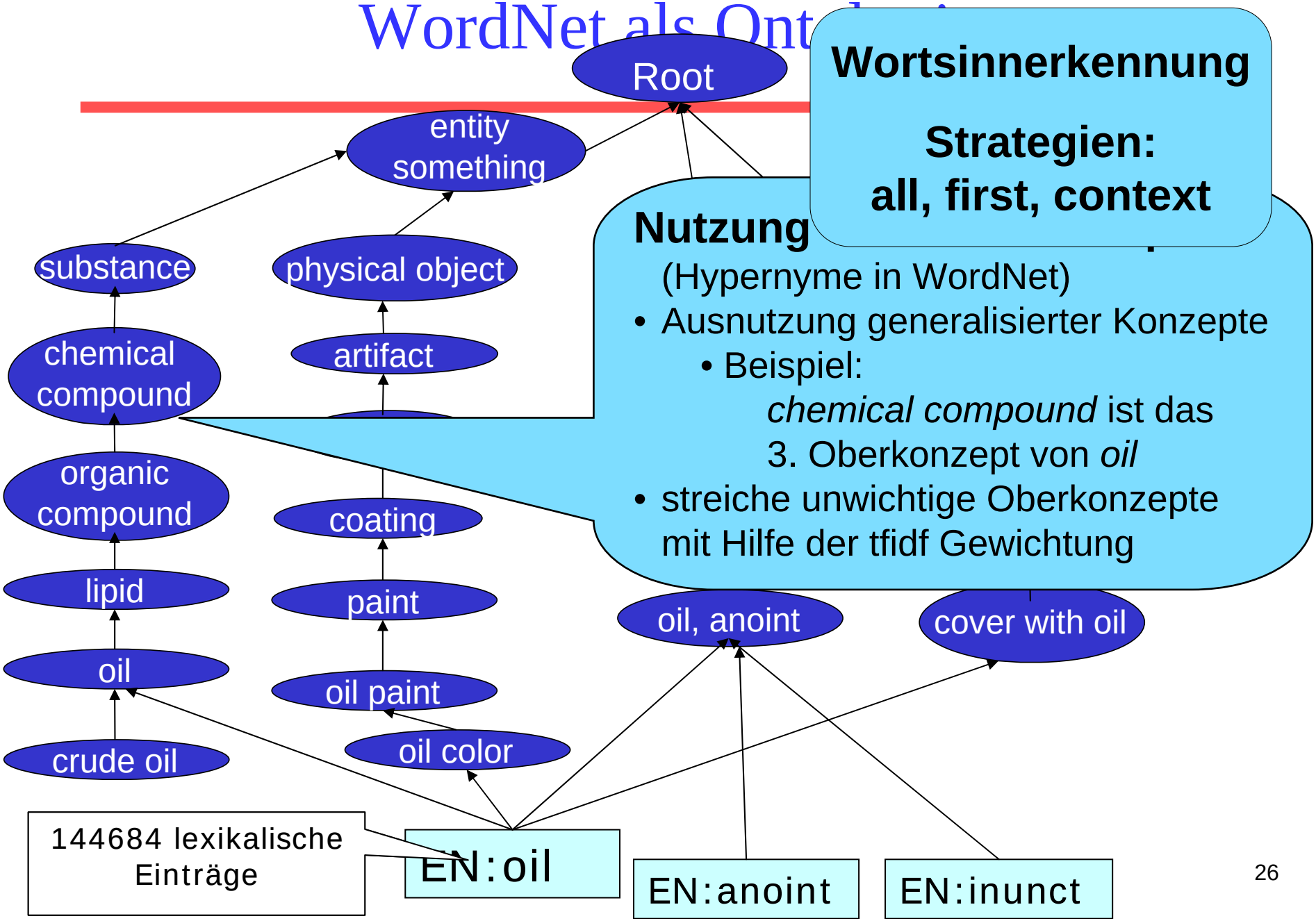
Oman	2
has	1
granted	1
term	1
crude	1
oil	2
customers	1
retroactive	1
discounts	1
...	...

Ontologie

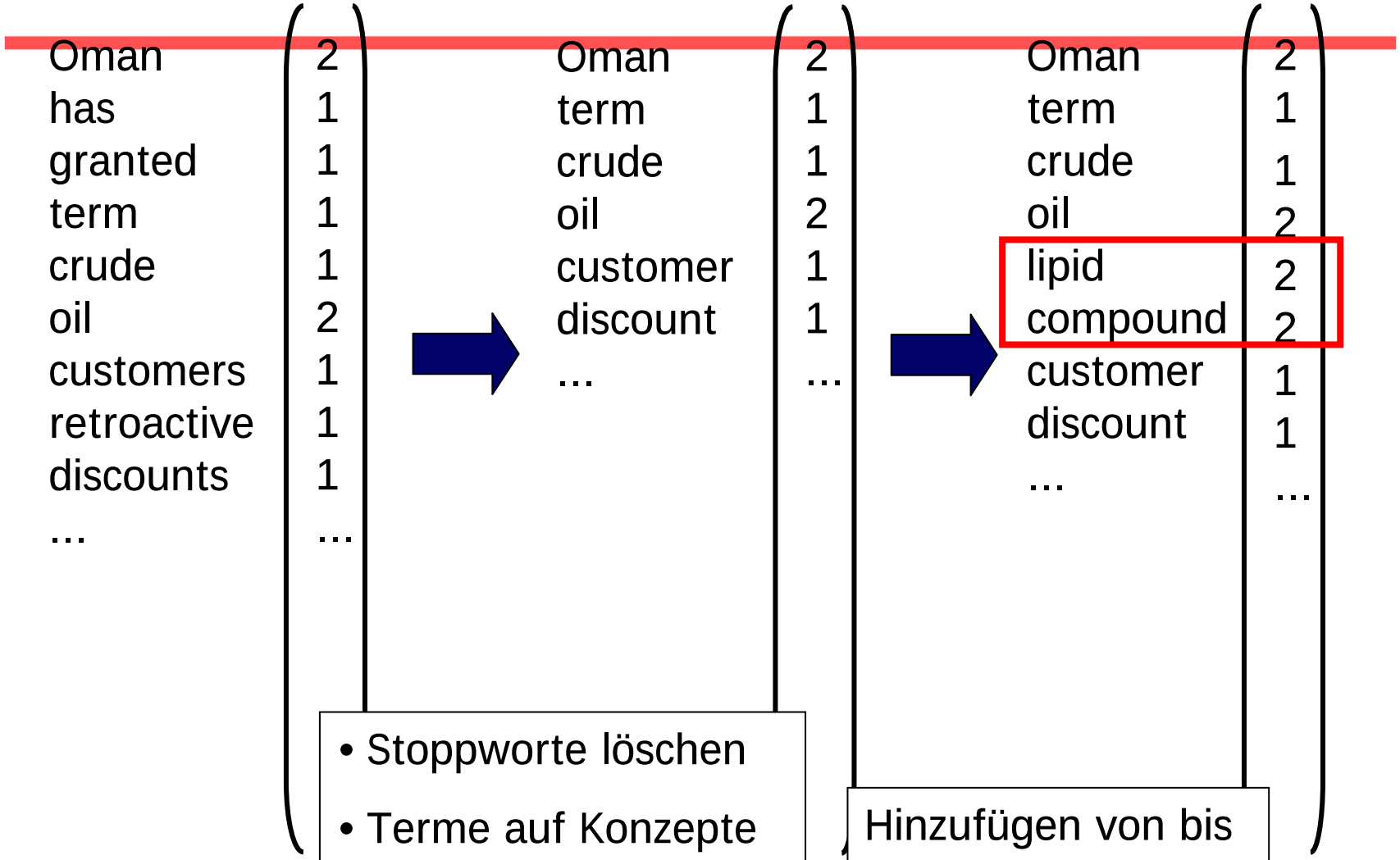
- Ontologie *O* repräsentiert das Hintergrundwissen
- Kernontologie besteht aus:
  - Menge von Konzepten: C
  - Konzepthierarchie oder Taxonomie: $\leq_C$
  - Lexikon: *Lex*



WordNet als Ont



Ontologiebasierte Repräsentation



Evaluierung Text-Clustern

Vergleich einer Clusterung $\mathcal{P}$ der Dokumentmenge D mit einer Referenzclusterung $\mathcal{L}$:

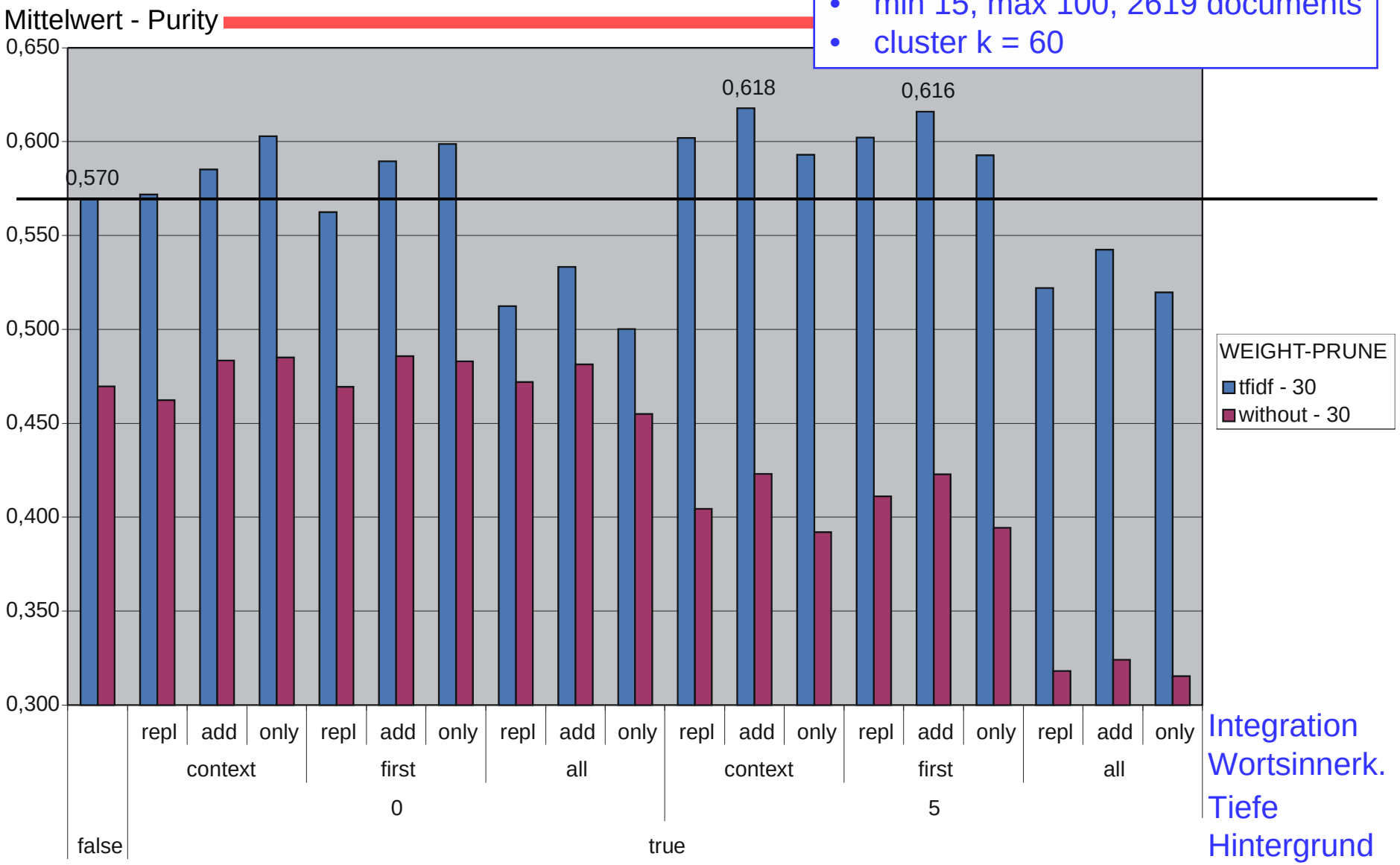
$$\text{Precision}(P, L) := \frac{|P \cap L|}{|P|}$$

$$\text{Purity}(\mathcal{P}, \mathcal{L}) := \sum_{P \in \mathcal{P}} \frac{|P|}{|D|} \max_{L \in \mathcal{L}} \text{Precision}(P, L)$$

$$\text{InversePurity}(\mathcal{P}, \mathcal{L}) := \sum_{L \in \mathcal{L}} \frac{|L|}{|D|} \max_{P \in \mathcal{P}} \text{Precision}(L, P)$$

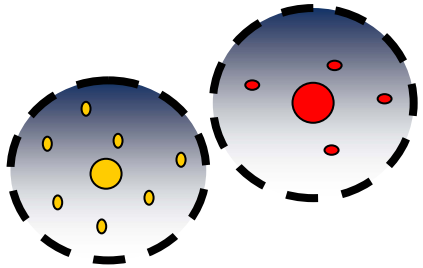
Evaluierung Text-Clustern

Evaluation parameter
- min 15, max 100, 2619 documents
- cluster k = 60



Erklärung durch Formale Begriffsanalyse

1. Repräsentiere jeden Cluster über die häufigsten Komponenten seines Mittelpunktvektors



2. Berechne Verbandsstruktur

(Hinweis: Größe des Verbandes kann exponentiell zur Anzahl der Objekte sein. Gut machbar für 50 oder 100 Cluster, aber nicht effizient für 10000 Dokumente)

Oman	1.1
granted	1.1
term	1.2
crude	1.1
oil	2.9
lipid	4.5
compound	2.3
customers	1.5
retroactive	0.2
discounts	0.3
...	

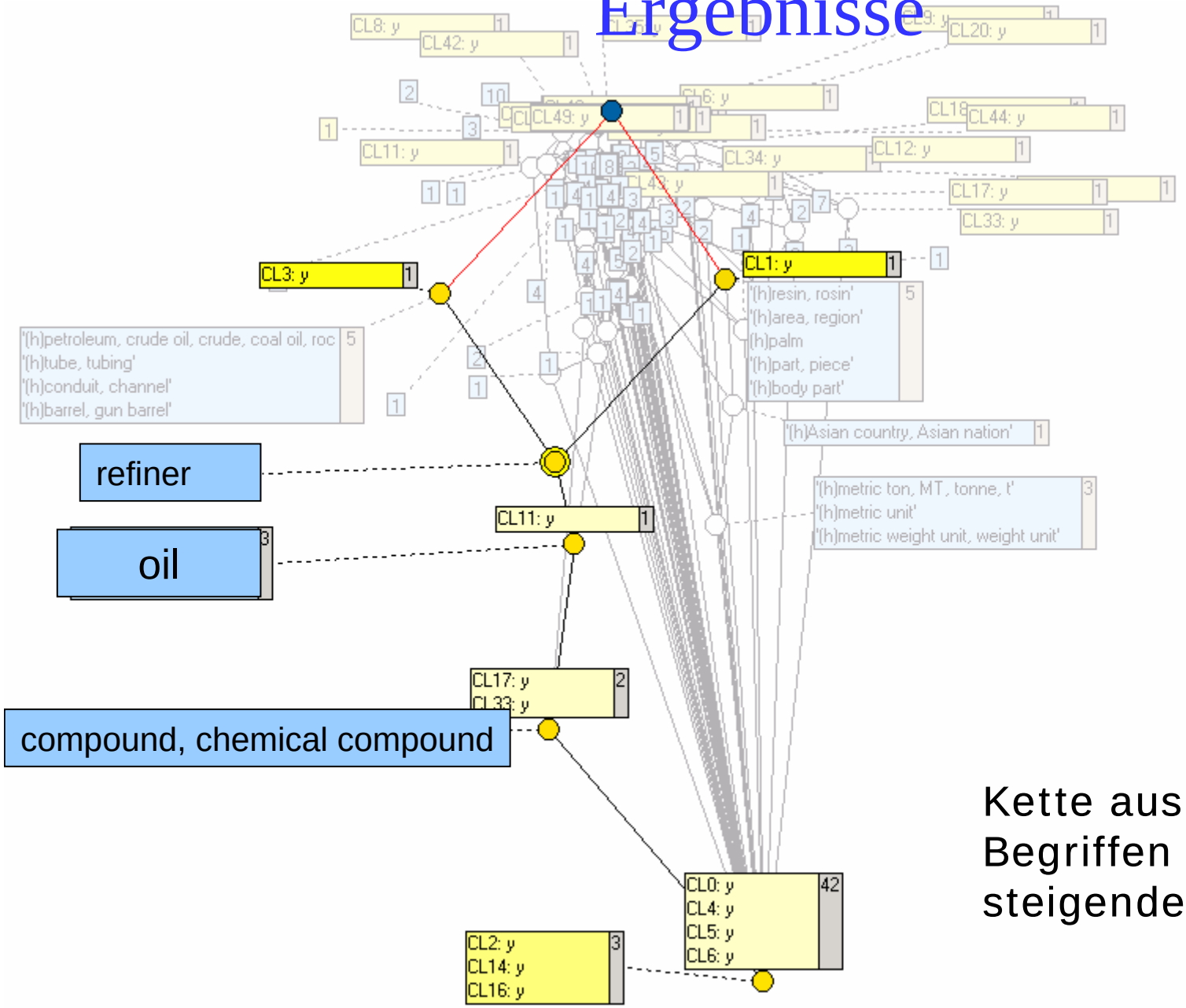
Oman	n
oil	y
lipid	y
compound	y
discounts	n

In jedem Cluster-zentroid werden alle Konzepte mit einem Wert kleiner als θ gelöscht (θ = 25% des max. Wertes)

Extrahierte Wort-/Konzeptliste

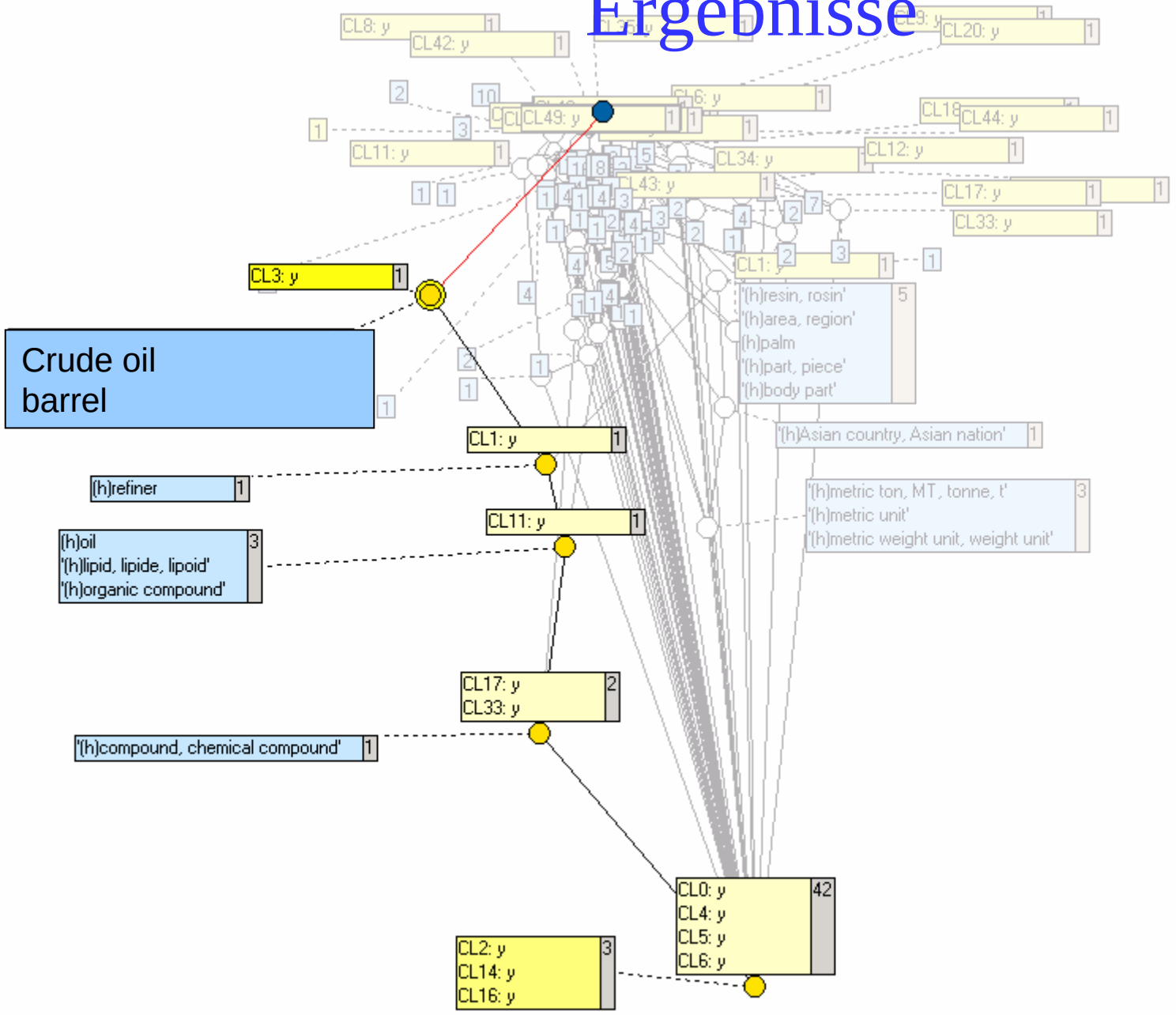
	Cluster 3		Cluster 9	
4	Irani, Iranian, Persian'	0,14	metric weight unit, weight	0,15
3	Iran, Islamic Republic of	0,13	metric ton, MT, tonne, t'	0,15
2	gulf	0,13	mass unit'	0,14
3	vessel, watercraft'	0,12	palm, thenar'	0,14
4	ship	0,12	area, region'	0,12
3	craft	0,12	unit of measurement, unit	0,10
2	Asian, Asiatic'	0,11	organic compound'	0,10
9	person of color, person of	0,10	oil	0,10
9	Asian country, Asian nati	0,10	lipid, lipide, lipoid'	0,10
3	oil tanker, oiler, tanker, ta	0,10	compound, chemical com	0,08

Ergebnisse

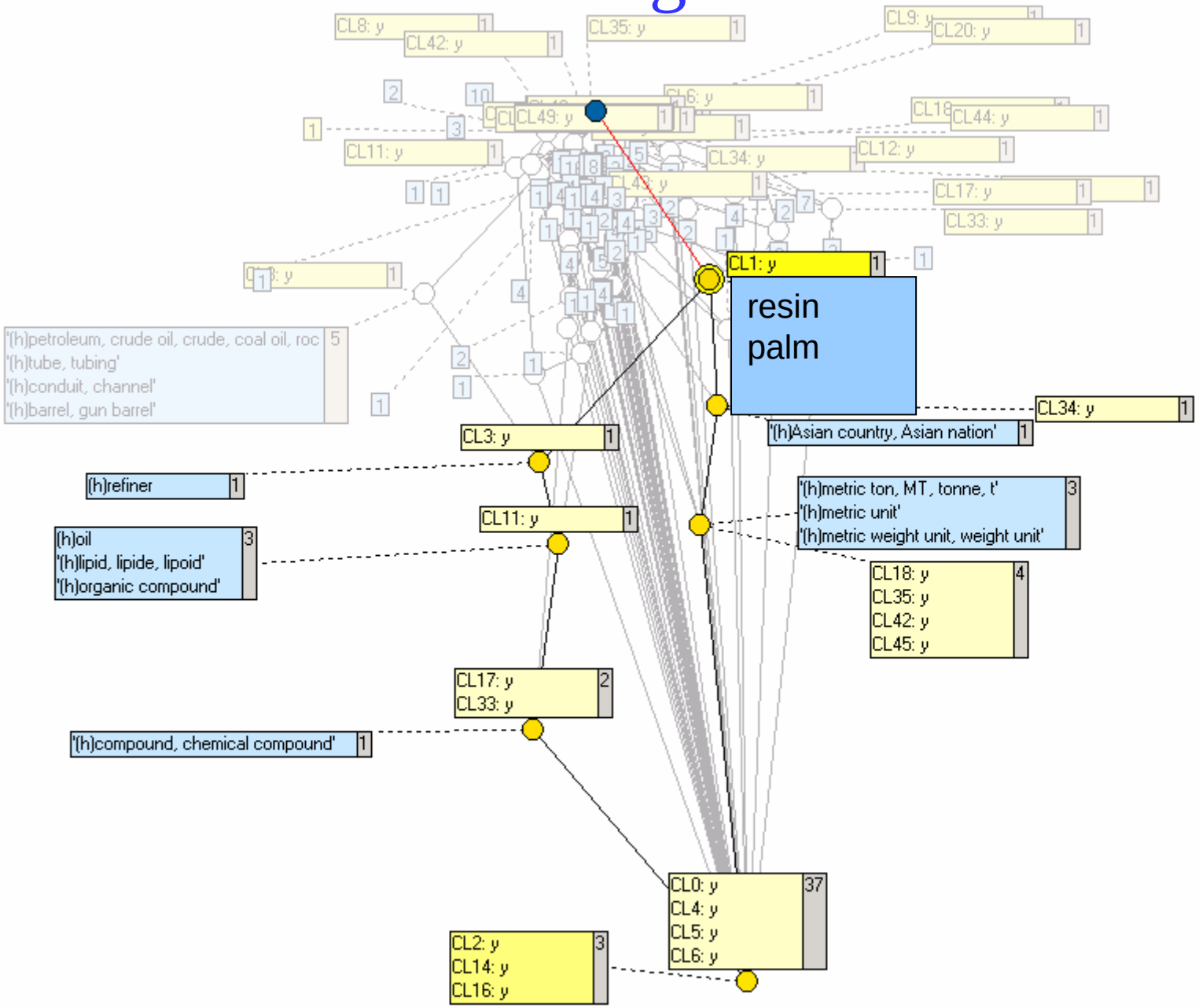


Kette aus formalen Begriffen mit steigender Spezifizität

Ergebnisse


Crude oil
barrel
CL3: y 1
CL1: y 1
(h)refiner 1
CL11: y 1
(h)oil
'(h)lipid, lipide, lipoid'
'(h)organic compound' 3
CL17: y
CL33: y 2
'(h)compound, chemical compound' 1
CL0: y
CL4: y
CL5: y
CL6: y 42
CL2: y
CL14: y
CL16: y 3
'(h)resin, rosin'
'(h)area, region'
(h)palm
'(h)part, piece'
'(h)body part' 5
'(h)Asian country, Asian nation' 1
'(h)metric ton, MT, tonne, t'
'(h)metric unit'
'(h)metric weight unit, weight unit' 3

Ergebnisse


CL1: y 1
resin
palm
'(h)petroleum, crude oil, crude, coal oil, roc
'(h)tube, tubing'
'(h)conduit, channel'
'(h)barrel, gun barrel' 5
CL3: y 1
(h)refiner 1
CL11: y 1
(h)oil
'(h)lipid, lipide, lipoid'
'(h)organic compound' 3
CL17: y
CL33: y 2
'(h)compound, chemical compound' 1
CL2: y
CL14: y
CL16: y 3
CL0: y
CL4: y
CL5: y
CL6: y 37
CL34: y 1
'(h)Asian country, Asian nation' 1
'(h)metric ton, MT, tonne, t'
'(h)metric unit'
'(h)metric weight unit, weight unit' 3
CL18: y
CL35: y
CL42: y
CL45: y 4