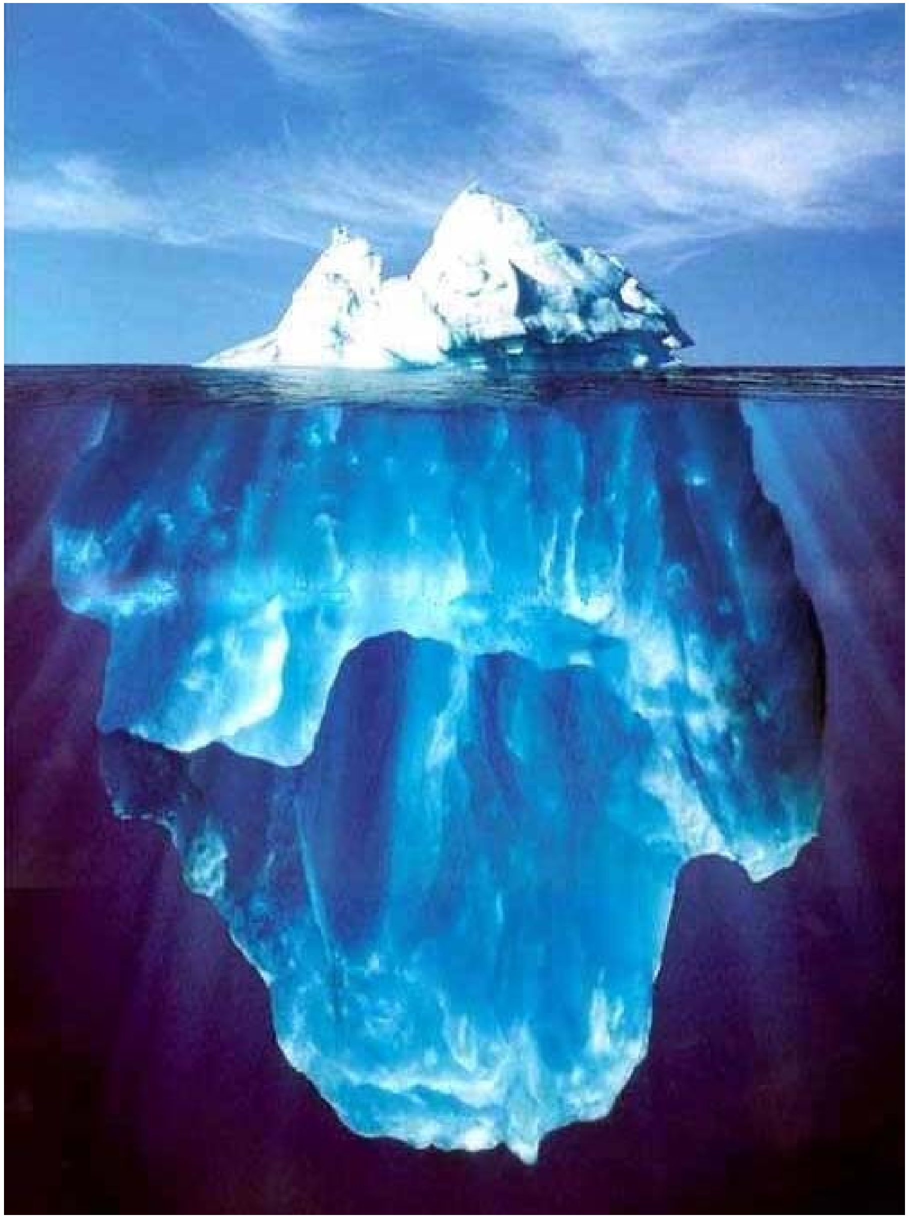


Effizientes Data Mining mit Formaler Begriffsanalyse

Vorlesung Knowledge Discovery

Kap. 9

UNI KASSEL VERSITÄT

FACHBEREICH MATHEMATIK / INFORMATIK

Fachgebiet Wissensverarbeitung

STIFTUNGSPROFESSUR DER GEMEINNÜTZIGEN HERTIE-STIFTUNG

1. **Motivation: Structuring the Frequent Itemset Space**

2. Formal Concept Analysis

3. Conceptual Clustering with Iceberg Concept Lattices

4. FCA-Based Mining of Association Rules

5. Text Clustering with Background Knowledge

Association Rules in a Nutshell

Association Rules are a popular data mining technique, e.g. for warehouse basket analysis: „Which items are frequently bought together?“

Toy Example:
Which activities can be frequently performed together in National Parks in California?

National Parks in California	NPS Guided Tours	Hiking	Horseback Riding	Swimming	Boating	Fishing	Bicycle Trail	Cross Country Trail
Cabrillo Natl. Mon.						×	×	
Channel Islands Natl. Park		×		×		×		
Death Valley Natl. Mon.	×	×	×	×			×	
Devils Postpile Natl. Mon.	×	×	×	×		×		
Fort Point Natl. Historic Site	×					×		
Golden Gate Natl. Recreation Area	×	×	×	×		×	×	
John Muir Natl. Historic Site	×							
Joshua Tree Natl. Mon.	×	×	×					
Kings Canyon Natl. Park	×	×	×			×		×
Lassen Volcanic Natl. Park	×	×	×	×	×	×		×
Lava Beds Natl. Mon.	×	×						
Muir Woods Natl. Mon.		×						
Pinnacles Natl. Mon.		×						
Point Reyes Natl. Seashore	×	×	×	×		×	×	
Redwood Natl. Park	×	×	×	×		×		
Santa Monica Mts. Natl. Recr. Area	×	×	×	×	×	×		
Sequoia Natl. Park	×	×	×			×		×
Whiskeytown-Shasta-Trinity Natl. Recr. Area	×	×	×	×	×	×		
Yosemite Natl. Park	×	×	×	×	×	×	×	×

{Swimming} $\rightarrow$ {Hiking}

conf = 100 %, supp = 10/19

#(swimming+hiking parks) / #(swimming parks)

#(swimming+hiking parks) / #(all parks)

Observation:

The rules

{ Boating } $\rightarrow$ { Hiking, NPS Guided Tours, Fishing }

{ Boating, Swimming } $\rightarrow$ { Hiking, NPS Guided Tours, Fishing }

have the same support and the same confidence, because the two sets

{ Boating } and { Boating, Swimming }

describe exactly the same set of parks.

	NPS Guided Tours	Hiking	Horseback Riding	Swimming	Boating	Fishing	Bicycle Trail	Cross Country Trail
Cabrillo Natl. Mon.						×	×	
Channel Islands Natl. Park		×		×		×		
Death Valley Natl. Mon.	×	×	×	×			×	
Devils Postpile Natl. Mon.	×	×	×	×		×		
Fort Point Natl. Historic Site	×					×		
Golden Gate Natl. Recreation Area	×	×	×	×		×	×	
John Muir Natl. Historic Site	×							
Joshua Tree Natl. Mon.	×	×	×					
Kings Canyon Natl. Park	×	×	×			×		×
Lassen Volcanic Natl. Park	×	×	×	×	×	×		×
Lava Beds Natl. Mon.	×	×						
Muir Woods Natl. Mon.		×						
Pinnacles Natl. Mon.		×						
Point Reyes Natl. Seashore	×	×	×	×		×	×	
Redwood Natl. Park	×	×	×	×		×		
Santa Monica Mts. Natl. Recr. Area	×	×	×	×	×	×		
Sequoia Natl. Park	×	×	×			×		×
Whiskeytown-Shasta-Trinity Natl. Recr. Area	×	×	×	×	×	×		
Yosemite Natl. Park	×	×	×	×	×	×	×	×

Conclusion:

It is sufficient to look at one of those sets!

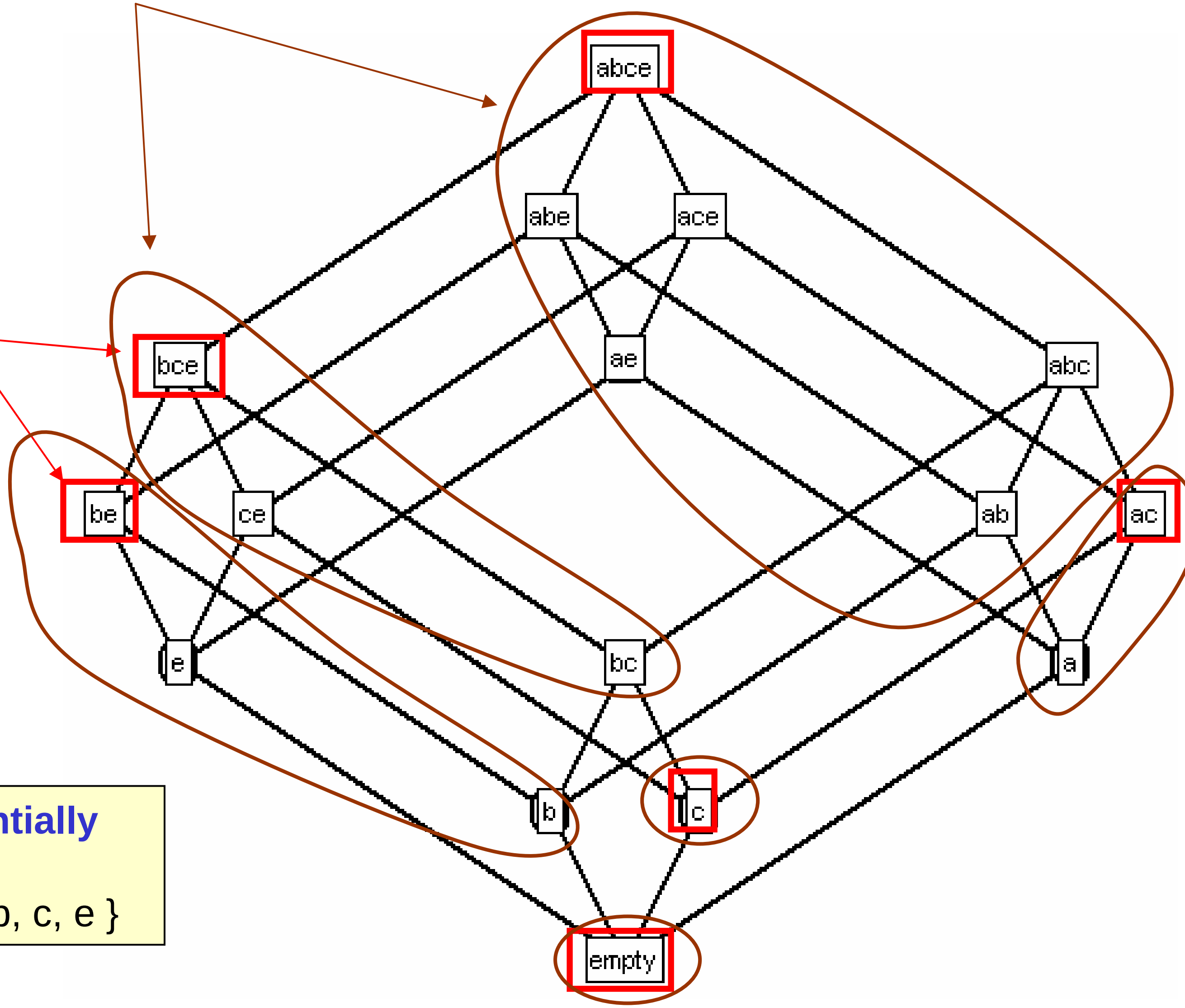
$\rightarrow$ **faster computation**

$\rightarrow$ **no redundant rules**

Another Toy Example:

	a	b	c	e
1	X		X	
2		X		X
3		X	X	X

Classes of itemsets describing the same sets of objects



Unique represen-
tatives of each class:

the **closed** itemsets

(or **concept intents**).

(6 instead of 16)

The **space of (potentially frequent) itemsets**:
the powerset of { a, b, c, e }

Bases of Association Rules

Classical Data Mining Task:
Find, for given minsupp, minconf $\in$ [0,1], all rules with support and confidence above these thresholds.

Two-Step Approach:

1. Compute all frequent itemsets (e.g., Apriori).
2. For each frequent itemset X and all its subsets Y:
   check $X \to Y$.

Our task:
Find a **basis** of rules, i.e., a minimal set of rules out of which all other rules can be derived.

Two-Step Approach:

1. Compute all frequent **closed** itemsets.
2. For each frequent **closed** itemset X and all its **closed** subsets Y:
   check $X \to Y$.

Association Rules and Formal Concept Analysis

Based on **Formal Concept Analysis (FCA).**

This relationship was discovered independently in 1998/9 at

- Clermont-Ferrand (Lakhal)
- Darmstadt (Stumme)
- New York (Zaki)

with Clermont being the fastest group developing algorithms (Close).

Our task:
Find a **basis** of rules, i.e., a minimal set of rules out of which all other rules can be derived.

Two-Step Approach:

1. Compute all frequent **closed** itemsets.
2. For each frequent **closed** itemset X and all its **closed** subsets Y:
   check $X \rightarrow Y$.

Association Rules and Formal Concept Analysis

Based on **Formal Concept Analysis (FCA).**

This relationship was discovered independently in 1998/9 at

- Clermont-Ferrand (Lakhal)
- Darmstadt (Stumme)
- New York (Zaki)

with Clermont being the fastest group developing algorithms (Close).

Our task:
Find a **basis** of rules, i.e., a minimal set of rules out of which all other rules can be derived.

Two-Step Approach:

1. Compute all frequent **closed** itemsets.
2. For each frequent **closed** itemset X and all its **closed** subsets Y:
   check $X \to Y$.

Structure of the Talk:

- Introduction to FCA
- Conceptual Clustering with FCA
- Mining Association Rules with FCA
- Frequent (Closed) Datalog Queries

1. Motivation: Structuring the Frequent Itemset Space

2. **Formal Concept Analysis**

3. Conceptual Clustering with Iceberg Concept Lattices

4. FCA-Based Mining of Association Rules

5. Text Clustering with Background Knowledge

Formal Concept Analysis

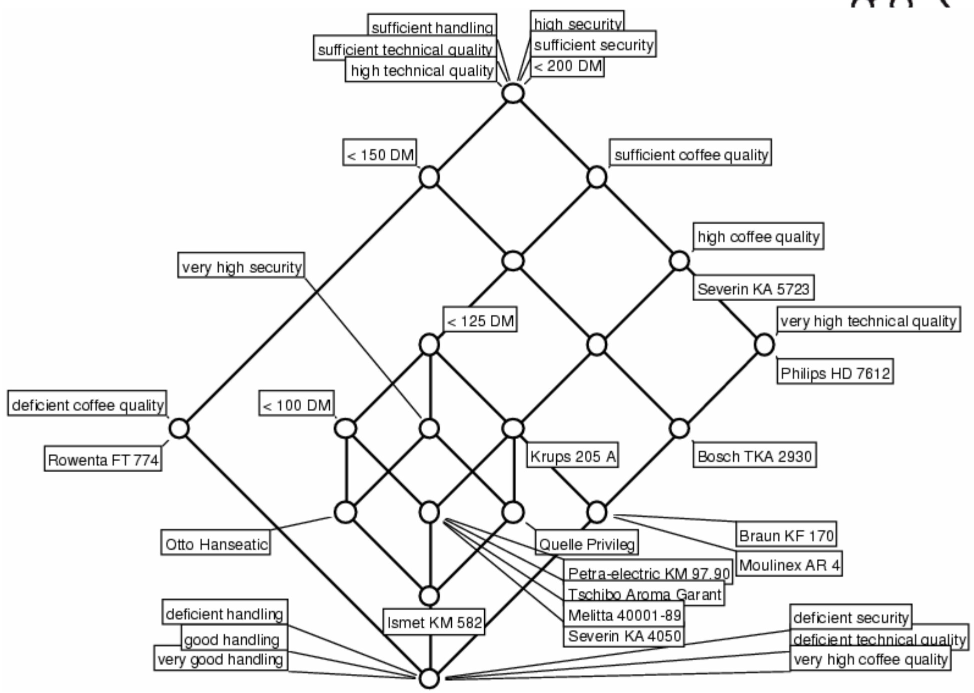
arose around 1980 in Darmstadt as a mathematical theory, which formalizes the concept of ‚concept‘.

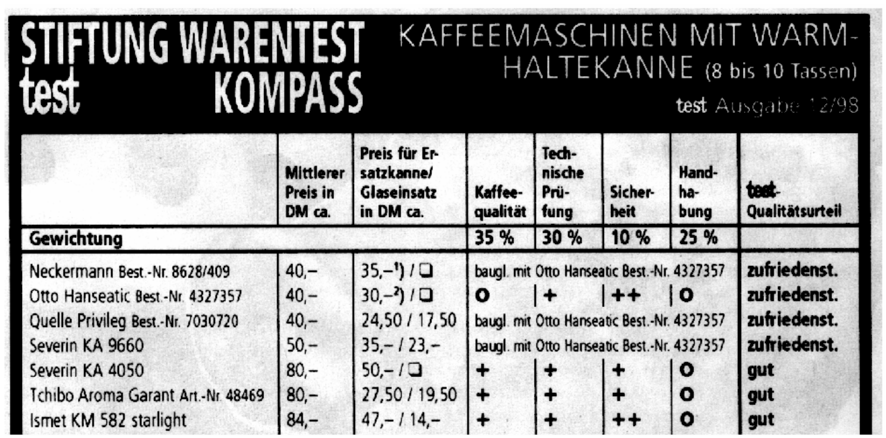
Since then, FCA has found many uses in Informatics, e.g. for

- Data Analysis,
- Information Retrieval,
- Knowledge Discovery,
- Software Engineering.

Based on datasets, FCA derives concept hierarchies.

FCA allows to generate and visualize concept hierarchies.





STIFTUNG WARENTEST test KOMPASS

KAFFEEMASCHINEN MIT WARMHALTEKANNE (8 bis 10 Tassen)

test Ausgabe 12/98

	Mittlerer Preis in DM ca.	Preis für Ersatzkanne/ Glaseinsatz in DM ca.	Kaffeequalität	Technische Prüfung	Sicherheit	Handhabung	test-Qualitätsurteil
Gewichtung			35 %	30 %	10 %	25 %	
Neckermann Best.-Nr. 8628/409	40,–	35,–[1] / ❏	baugl. mit Otto Hanseatic Best.-Nr. 4327357				zufriedenst.
Otto Hanseatic Best.-Nr. 4327357	40,–	30,–[2] / ❏	o	+	++	o	zufriedenst.
Quelle Privileg Best.-Nr. 7030720	40,–	24,50 / 17,50	baugl. mit Otto Hanseatic Best.-Nr. 4327357				zufriedenst.
Severin KA 9660	50,–	35,– / 23,–	baugl. mit Otto Hanseatic Best.-Nr. 4327357				zufriedenst.
Severin KA 4050	80,–	50,– / ❏	+	+	+	o	gut
Tchibo Aroma Garant Art.-Nr. 48469	80,–	27,50 / 19,50	+	+	+	o	gut
Ismet KM 582 starlight	84,–	47,– / 14,–	+	+	++	o	gut

FCA models **concepts** as **units of thought**, consisting of two parts:

- The **extension** consists of all objects belonging to the concept.
- The **intension** consists of all attributes common to all those objects.

Some **typical applications**:

- database marketing
- email management system
- developing qualitative theories in music estethics
- analysis of flight movements at Frankfurt airport

Formal Concept Analysis

Def.: A **formal context** is a triple (*G*,*M*,*I*), where

- *G* is a set of objects,
- *M* is a set of attributes
- and *I* is a relation between *G* and *M*.

- (*g*,*m*)∈*I* is read as „object *g* has attribute *m*“.

National Parks in California	NPS Guided Tours	Hiking	Horseback Riding	Swimming	Boating	Fishing	Bicycle Trail	Cross Country Trail
Cabrillo Natl. Mon.						×	×	
Channel Islands Natl. Park		×		×		×		
Death Valley Natl. Mon.	×	×	×	×			×	
Devils Postpile Natl. Mon.	×	×	×	×		×		
Fort Point Natl. Historic Site	×					×		
Golden Gate Natl. Recreation Area	×	×	×	×		×	×	
John Muir Natl. Historic Site	×							
Joshua Tree Natl. Mon.	×	×	×					
Kings Canyon Natl. Park	×	×	×			×		×
Lassen Volcanic Natl. Park	×	×	×	×	×	×		×
Lava Beds Natl. Mon.	×	×						
Muir Woods Natl. Mon.		×						
Pinnacles Natl. Mon.		×						
Point Reyes Natl. Seashore	×	×	×	×		×	×	
Redwood Natl. Park	×	×	×	×		×		
Santa Monica Mts. Natl. Recr. Area	×	×	×	×	×	×		
Sequoia Natl. Park	×	×	×			×		×
Whiskeytown-Shasta-Trinity Natl. Recr. Area	×	×	×	×	×	×		
Yosemite Natl. Park	×	×	×	×	×	×	×	×

For $A \subseteq G$, we define

$A' := \{ m \in M \mid \forall g \in A: (g,m) \in I \}$.

For $B \subseteq M$, we define dually

$B' := \{ g \in G \mid \forall m \in B: (g,m) \in I \}$.

A´

National Parks in California	NPS Guided Tours	Hiking	Horseback Riding	Swimming	Boating	Fishing	Bicycle Trail	Cross Country Trail
Cabrillo Natl. Mon.						×	×	
Channel Islands Natl. Park		×		×		×		
Death Valley Natl. Mon.	×	×	×	×			×	
Devils Postpile Natl. Mon.	×	×	×	×		×		
Fort Point Natl. Historic Site	×					×		
Golden Gate Natl. Recreation Area	×	×	×	×		×	×	
John Muir Natl. Historic Site	×							
Joshua Tree Natl. Mon.	×	×	×					
Kings Canyon Natl. Park	×	×	×			×		×
Lassen Volcanic Natl. Park	×	×	×	×	×	×		×
Lava Beds Natl. Mon.	×	×						
Muir Woods Natl. Mon.		×						
Pinnacles Natl. Mon.		×						
Point Reyes Natl. Seashore	×	×	×	×		×	×	
Redwood Natl. Park	×	×	×	×		×		
Santa Monica Mts. Natl. Recr. Area	×	×	×	×	×	×		
Sequoia Natl. Park	×	×	×			×		×
Whiskeytown-Shasta-Trinity Natl. Recr. Area	×	×	×	×	×	×		
Yosemite Natl. Park	×	×	×	×	×	×	×	×

A

Def.: A **formal concept**

is a pair (*A*,*B*) where

- *A* is a set of objects (the **extent** of the concept),
- *B* is a set of attributes (the **intent** of the concept),
- A‘ = B and B‘ = A.

= closed itemset

Intent *B*

Extent *A*

National Parks in California	*NPS Guided Tours*	*Hiking*	*Horseback Riding*	*Swimming*	*Boating*	*Fishing*	*Bicycle Trail*	*Cross Country Trail*
Cabrillo Natl. Mon.						×	×	
Channel Islands Natl. Park		×		×		×		
Death Valley Natl. Mon.	×	×	×	×			×	
Devils Postpile Natl. Mon.	×	×	×	×		×		
Fort Point Natl. Historic Site	×					×		
Golden Gate Natl. Recreation Area	×	×	×	×		×	×	
John Muir Natl. Historic Site	×							
Joshua Tree Natl. Mon.	×	×	×					
Kings Canyon Natl. Park	×	×	×			×		×
Lassen Volcanic Natl. Park	×	×	×	×	×	×		×
Lava Beds Natl. Mon.	×	×						
Muir Woods Natl. Mon.		×						
Pinnacles Natl. Mon.		×						
Point Reyes Natl. Seashore	×	×	×	×		×	×	
Redwood Natl. Park	×	×	×	×		×		
Santa Monica Mts. Natl. Recr. Area	×	×	×	×	×	×		
Sequoia Natl. Park	×	×	×			×		×
Whiskeytown-Shasta-Trinity Natl. Recr. Area	×	×	×	×	×	×		
Yosemite Natl. Park	×	×	×	×	×	×	×	×

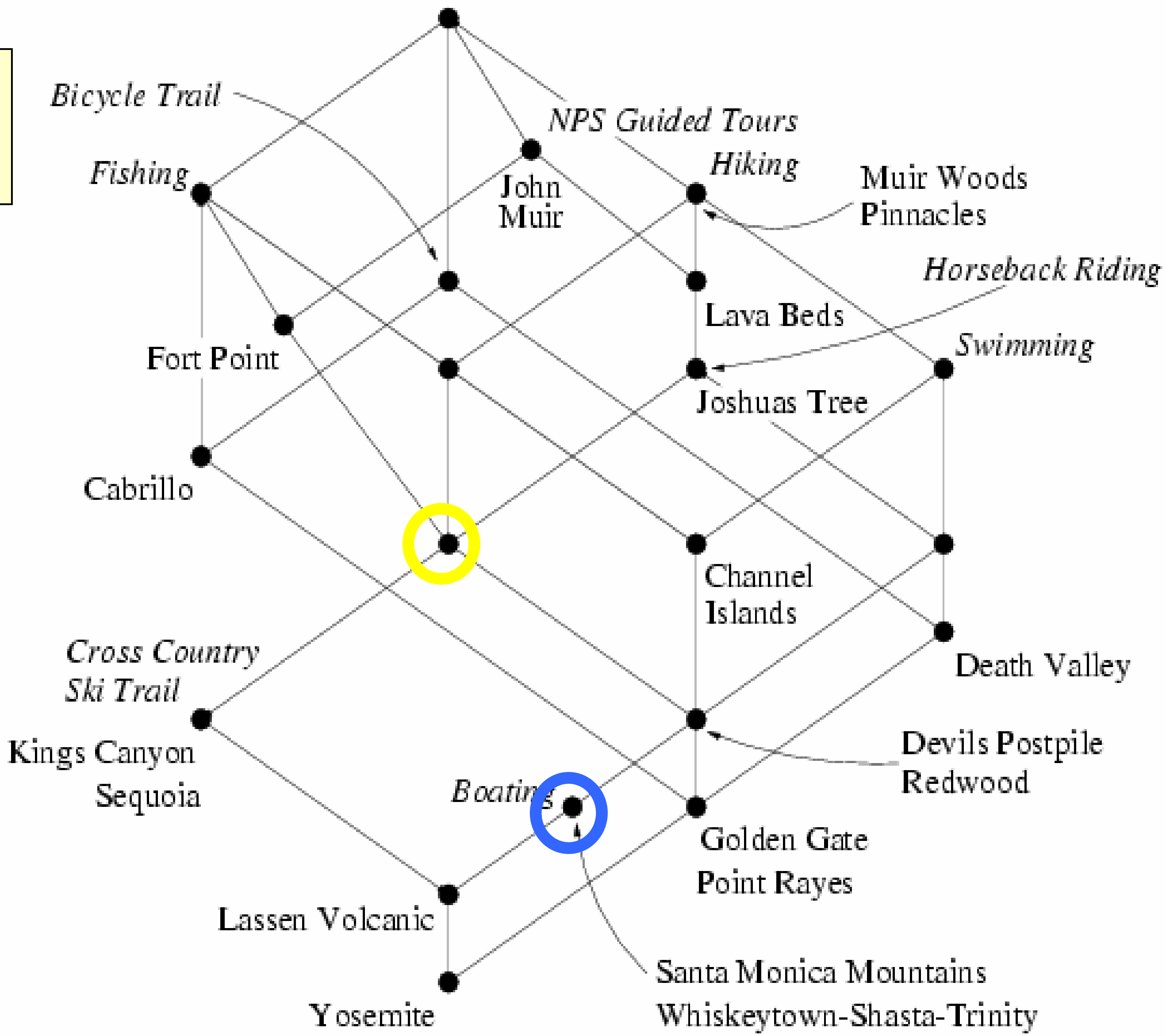
The blue concept is a **subconcept** of the yellow one, since its extent is contained in the yellow one.

(⇔ the yellow intent is contained in the blue one.)

National Parks in California	NPS Guided Tours	Hiking	Horseback Riding	Swimming	Boating	Fishing	Bicycle Trail	Cross Country Trail
Cabrillo Natl. Mon.						×	×	
Channel Islands Natl. Park		×		×		×		
Death Valley Natl. Mon.	×	×	×	×			×	
Devils Postpile Natl. Mon.	×	×	×	×		×		
Fort Point Natl. Historic Site	×					×		
Golden Gate Natl. Recreation Area	×	×	×	×		×	×	
John Muir Natl. Historic Site	×							
Joshua Tree Natl. Mon.	×	×	×					
Kings Canyon Natl. Park	×	×	×			×		×
Lassen Volcanic Natl. Park	×	×	×	×	×	×		×
Lava Beds Natl. Mon.	×	×						
Muir Woods Natl. Mon.		×						
Pinnacles Natl. Mon.		×						
Point Reyes Natl. Seashore	×	×	×	×		×	×	
Redwood Natl. Park	×	×	×	×		×		
Santa Monica Mts. Natl. Recr. Area	×	×	×	×	×	×		
Sequoia Natl. Park	×	×	×			×		×
Whiskeytown-Shasta-Trinity Natl. Recr. Area	×	×	×	×	×	×		
Yosemite Natl. Park	×	×	×	×	×	×	×	×

The **concept lattice** of the National Parks in California

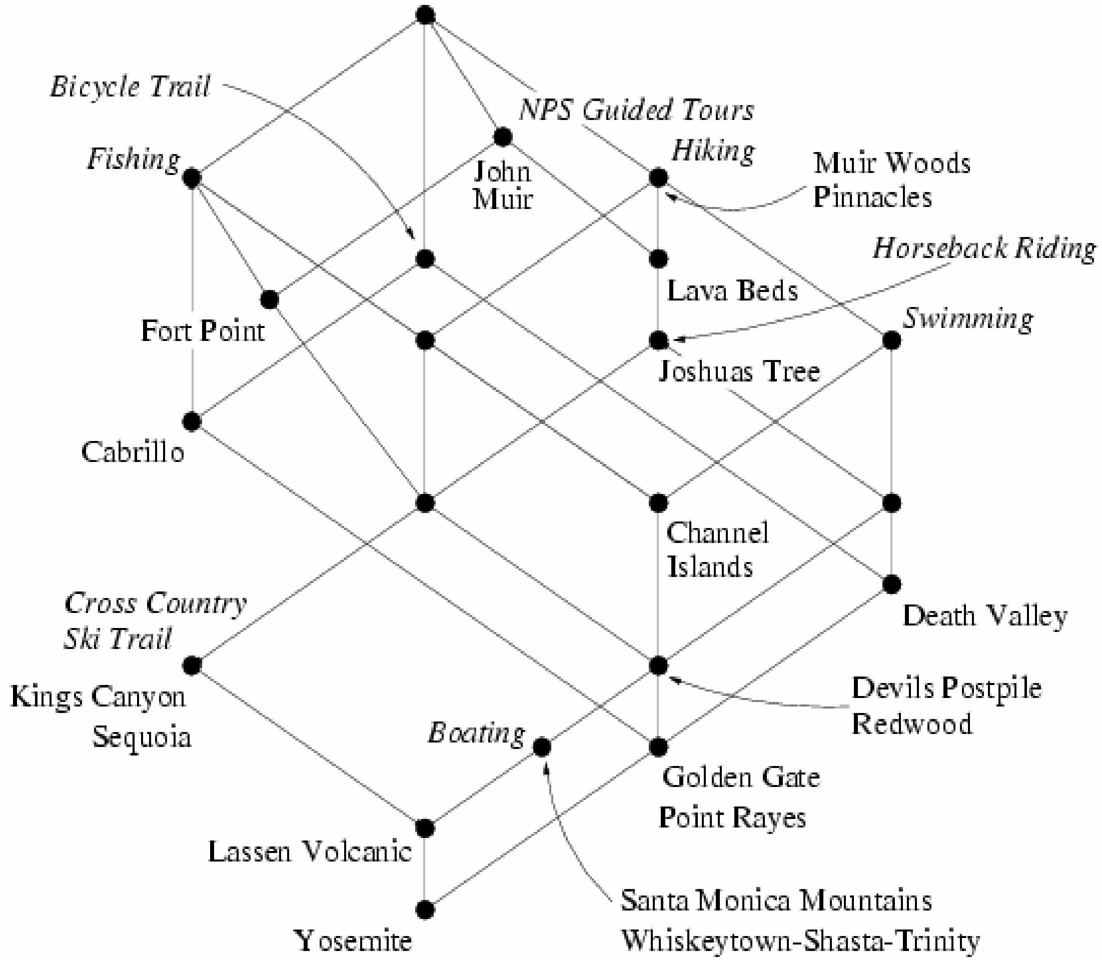
National Parks in California	NPS Guided Tours	Hiking	Horseback Riding	Swimming	Boating	Fishing	Bicycle Trail	Cross Country Trail
Cabrillo Natl. Mon.						×	×	
Channel Islands Natl. Park		×		×		×		
Death Valley Natl. Mon.	×	×	×	×			×	
Devils Postpile Natl. Mon.	×	×	×	×		×		
Fort Point Natl. Historic Site	×					×		
Golden Gate Natl. Recreation Area	×	×	×	×		×	×	
John Muir Natl. Historic Site	×							
Joshua Tree Natl. Mon.	×	×	×					
Kings Canyon Natl. Park	×	×	×			×		×
Lassen Volcanic Natl. Park	×	×	×	×	×	×		×
Lava Beds Natl. Mon.	×	×						
Muir Woods Natl. Mon.		×						
Pinnacles Natl. Mon.		×						
Point Reyes Natl. Seashore	×	×	×	×		×	×	
Redwood Natl. Park	×	×	×	×		×		
Santa Monica Mts. Natl. Recr. Area	×	×	×	×	×	×		
Sequoia Natl. Park	×	×	×			×		×
Whiskeytown-Shasta-Trinity Natl. Recr. Area	×	×	×	×	×	×		
Yosemite Natl. Park	×	×	×	×	×	×	×	×



Implications

Def.: An **implication**
$X \rightarrow Y$ holds in a context, if every object having all attributes in X also has all attributes in Y.

(= Association rule with 100% confidence)



- **Examples:**

  { Swimming } $\rightarrow$ { Hiking }

  { Boating } $\rightarrow$ { Swimming, Hiking, NPS Guided Tours, Fishing }

  { Bicycle Trail, NPS Guided Tours } $\rightarrow$ { Swimming, Hiking }

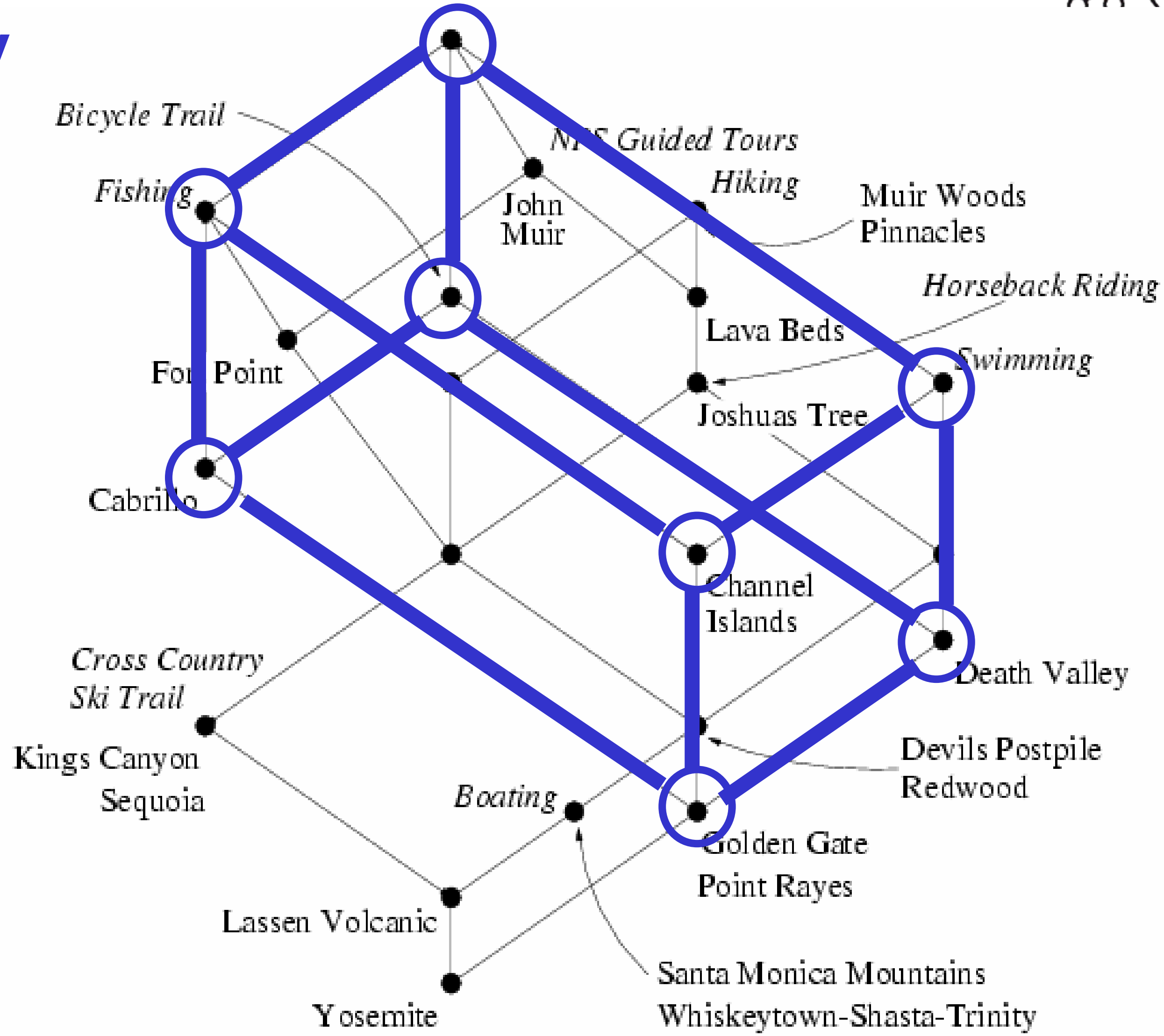
Independency

Attributes are independent if they span a hyper-cube (i.e., if all 2^n combi-nations occur).

Example:

- Fishing
- Bicycle Trail
- Swimming

are independent attributes.



1. Motivation: Structuring the Frequent Itemset Space

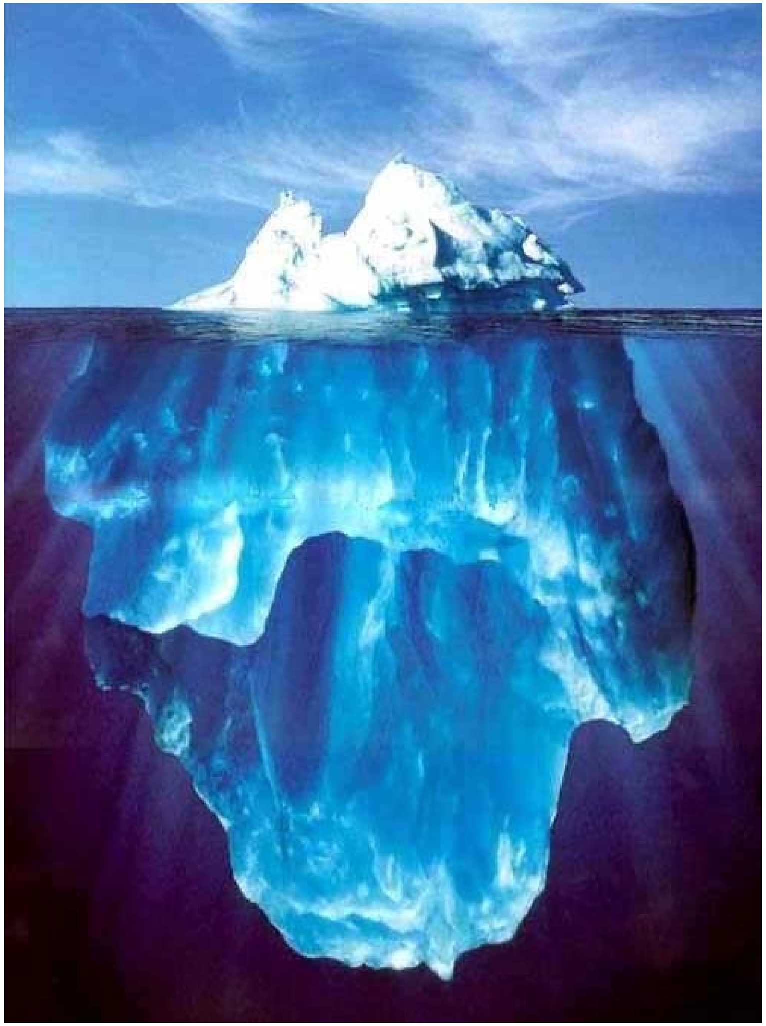
2. Formal Concept Analysis

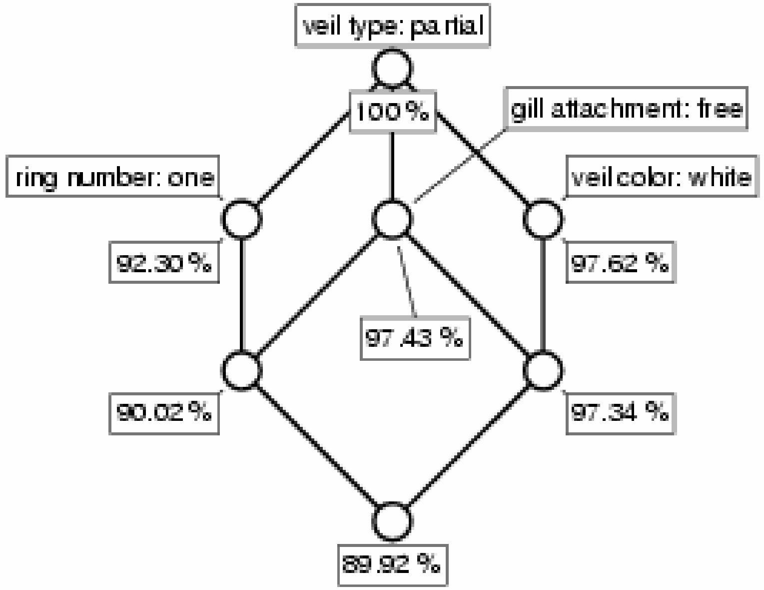
3. **Conceptual Clustering with Iceberg Concept Lattices**

4. FCA-Based Mining of Association Rules

5. Text Clustering with Background Knowledge

Iceberg Concept Lattices

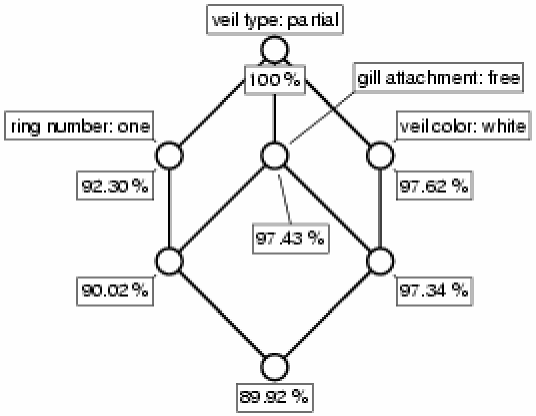




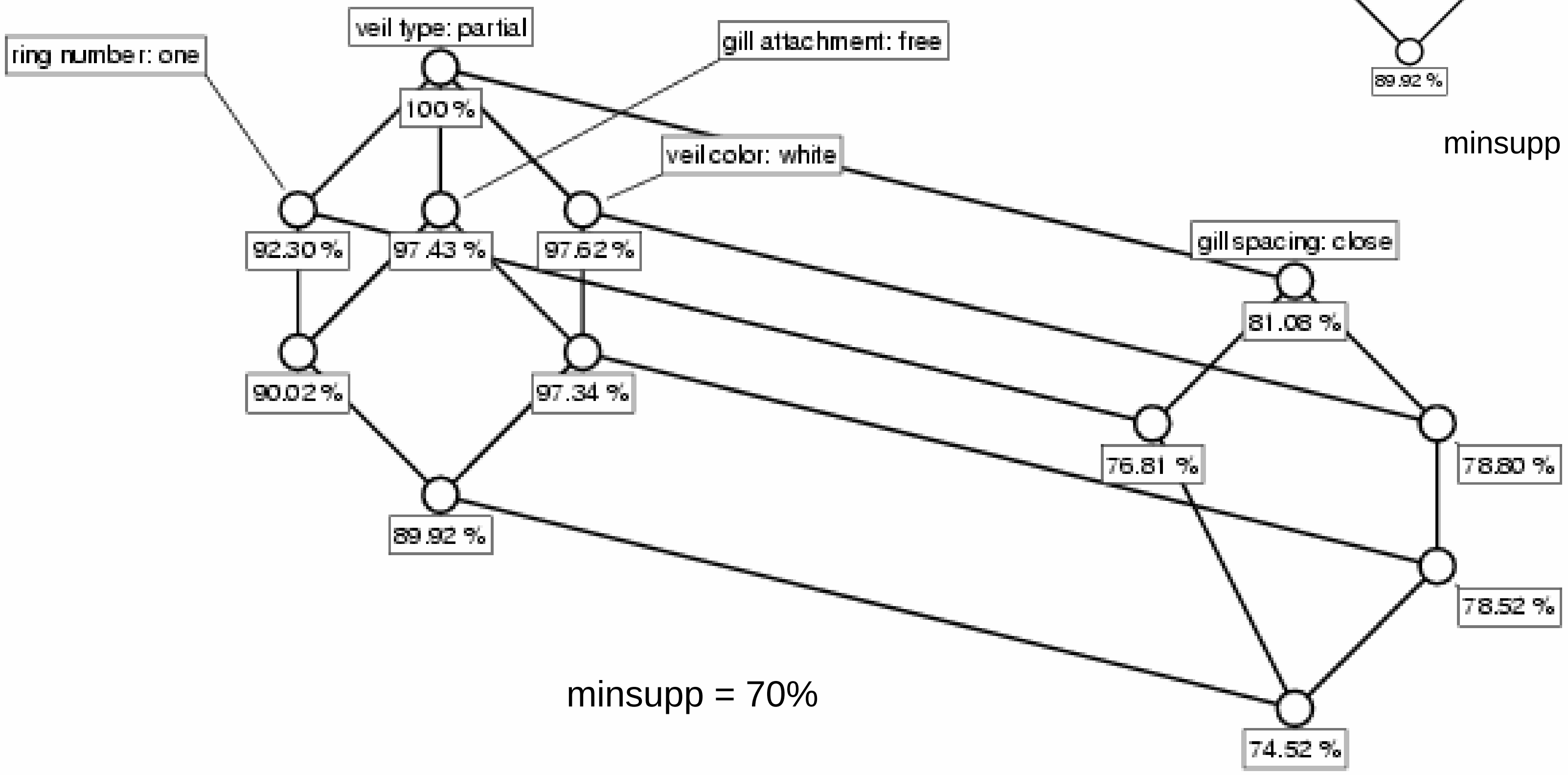
minsupp = 85%

For minsupp = 85% the seven most general of the 32.086 concepts of the Mushrooms database http:\\kdd.ics.uci.edu are shown.

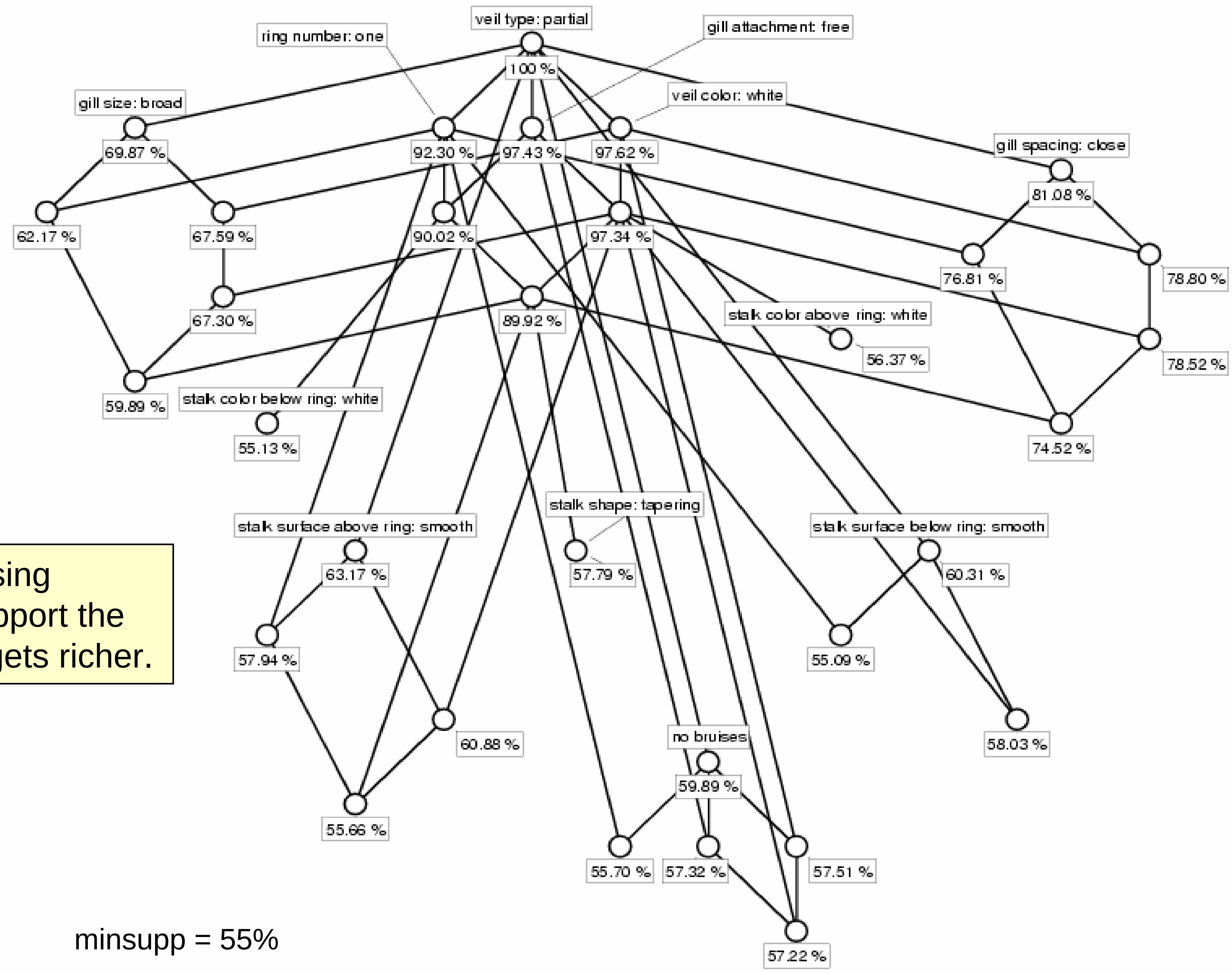
Iceberg Concept Lattices



minsupp = 85%



minsupp = 70%



With decreasing minimum support the information gets richer.

minsupp = 55%

Iceberg Concept Lattices and Frequent Itemsets

Iceberg concept lattices are a condensed representation of frequent itemsets:

supp(X) = supp(X‘‘)

minsupp	# frequent closed itemsets	# frequent itemsets
85 %	7	16
70 %	12	32
55 %	32	116
0 %	32.086	2^{80}

Difference between frequent concepts and frequent itemsets in the mushrooms database.

1. Motivation: Structuring the Frequent Itemset Space

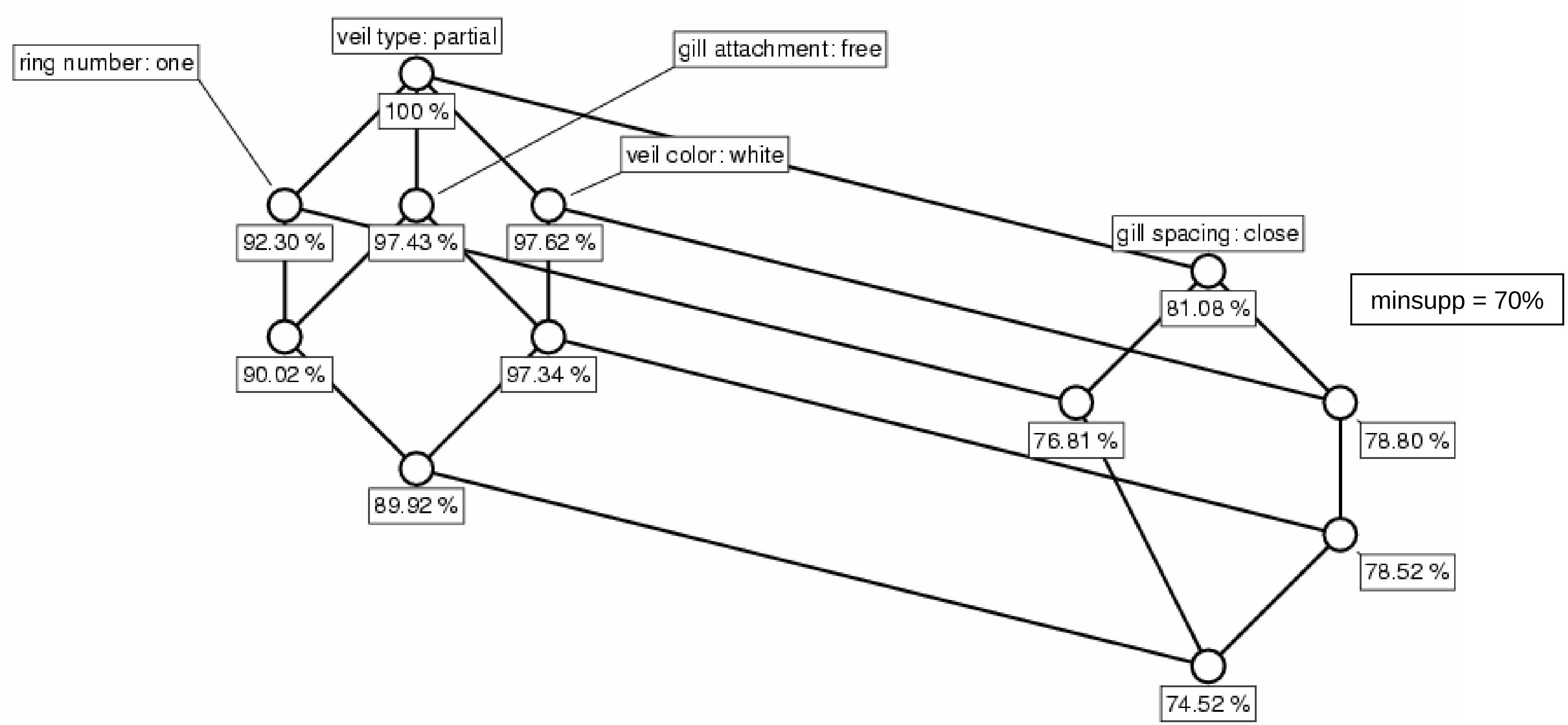
2. Formal Concept Analysis

3. Conceptual Clustering with Iceberg Concept Lattices

4. **FCA-Based Mining of Association Rules**

5. Text Clustering with Background Knowledge

Advantage of the use of iceberg concept lattices (compared to frequent itemsets)



32 frequent itemsets are represented by 12 frequent concept intents

→ **more efficient computation (e.g. TITANIC)**

→ **fewer rules (without information loss!)**

- From $\mathrm{supp}(B) = \mathrm{supp}\,(B'')$ follows:

Theorem: $X \to Y$ and $X'' \to Y''$ have the same support and the same confidence.

Hence for computing association rules, it is sufficient to compute the supports of all frequent sets with $B = B''$ (i.e., the intents of the iceberg concept lattice).

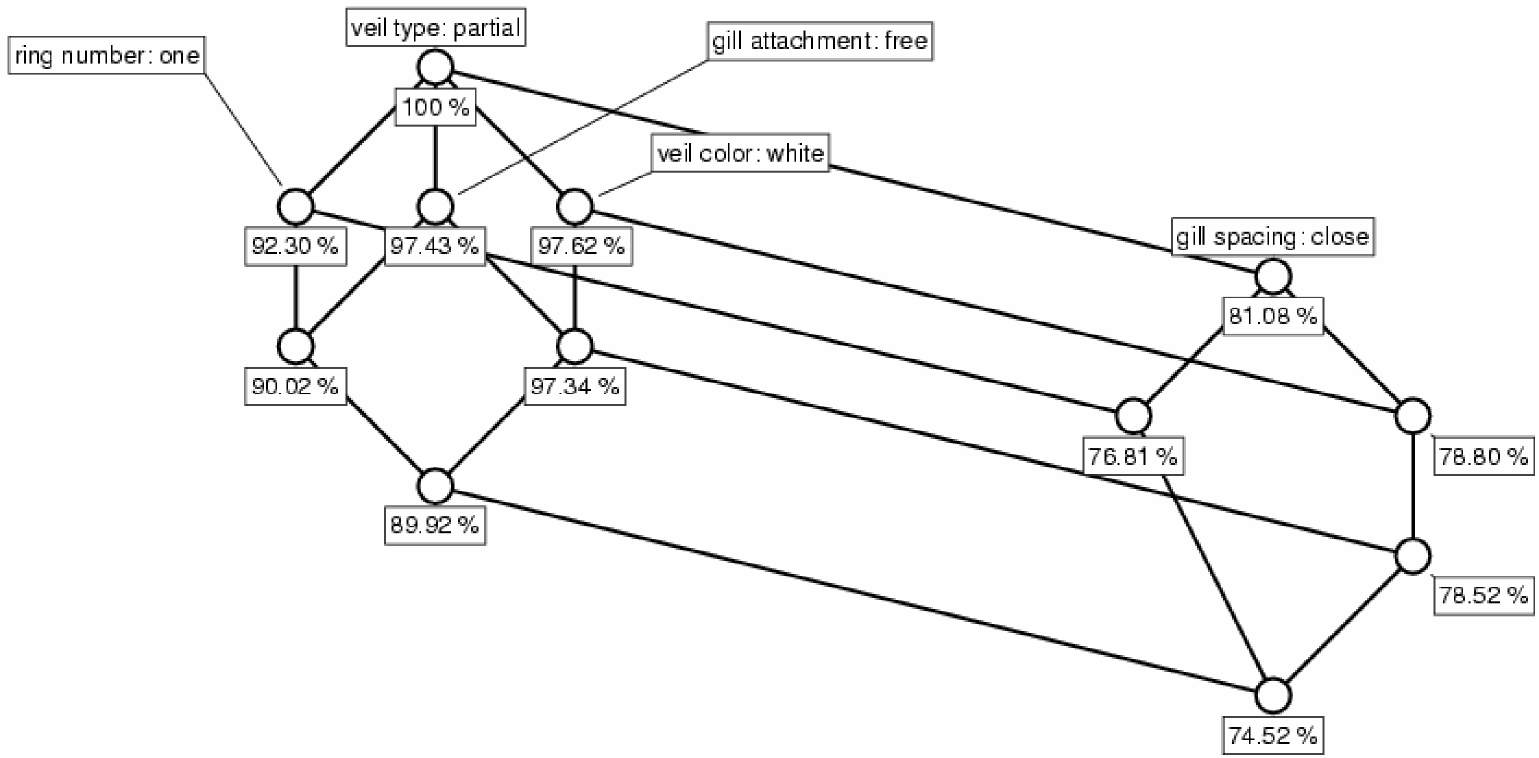
Association rules can be visualized in the iceberg concept lattice:

- **exact rules**
- **approximate rules**

conf = 100 %

conf < 100 %

Exact association rules



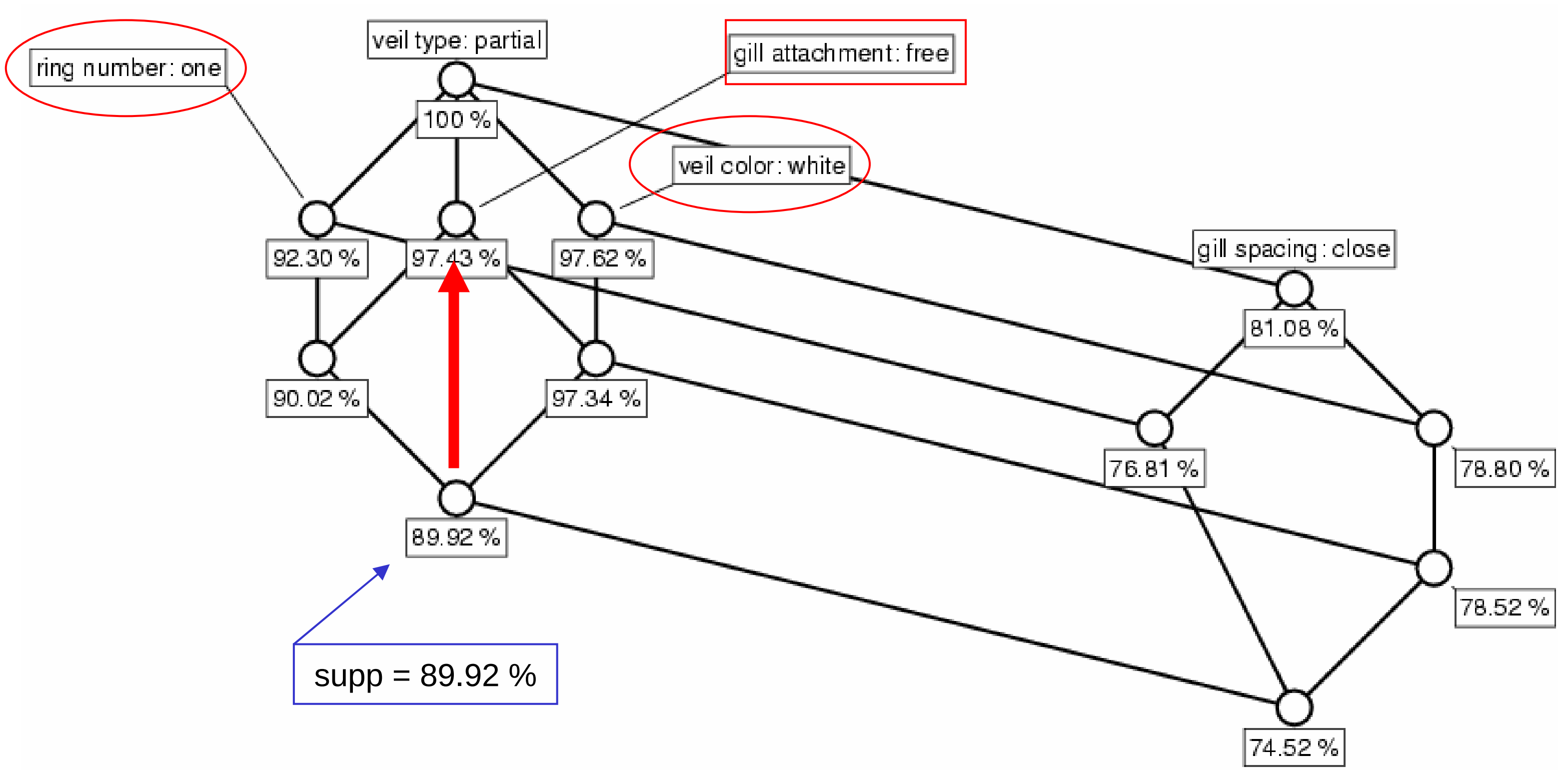
Association rules can be visualized in the iceberg concept lattice:

- **exact rules**
- **approximate rules**

conf = 100 %

conf < 100 %

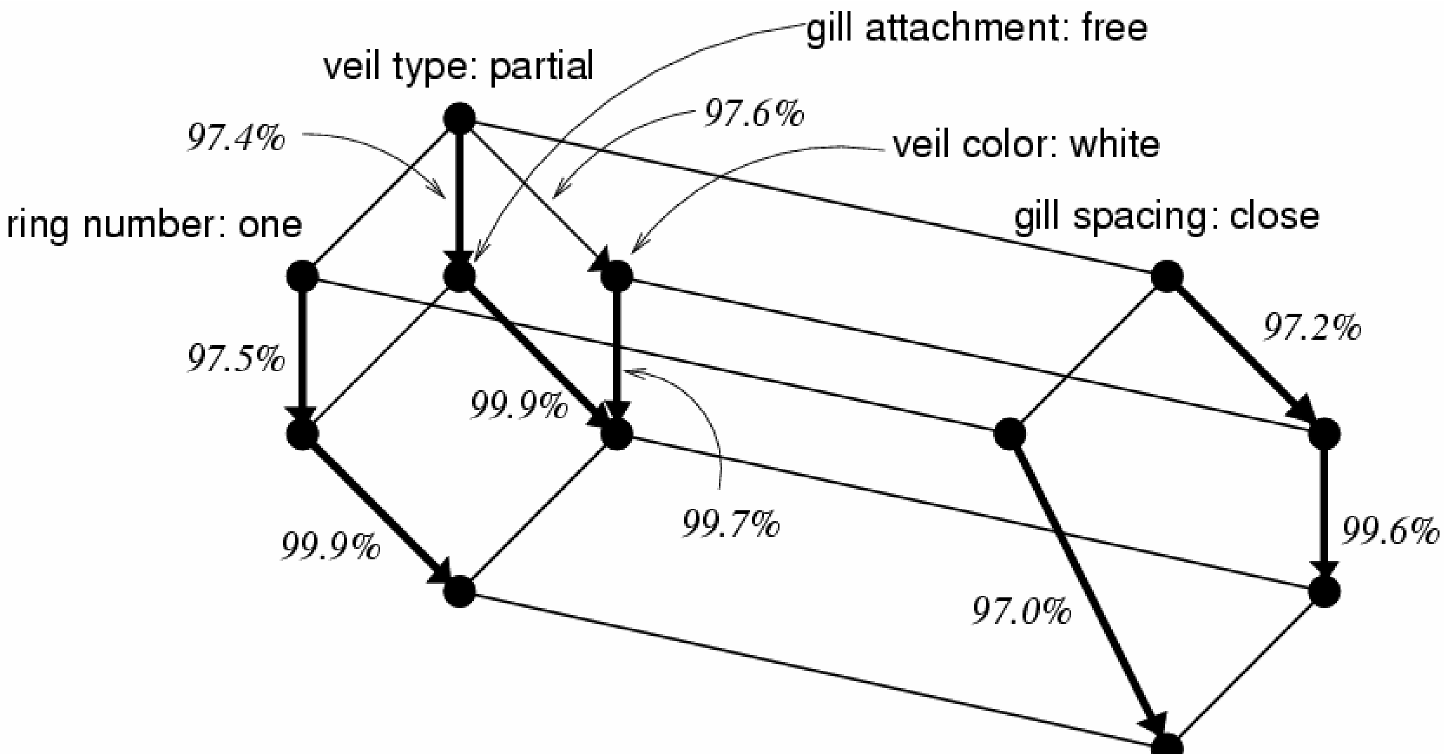
Exact association rules



{ring number: one, veil color: white} → {gill attachment: free}

supp = 89.92 % conf = 100 %.

Luxenburger Basis for approximate association rules



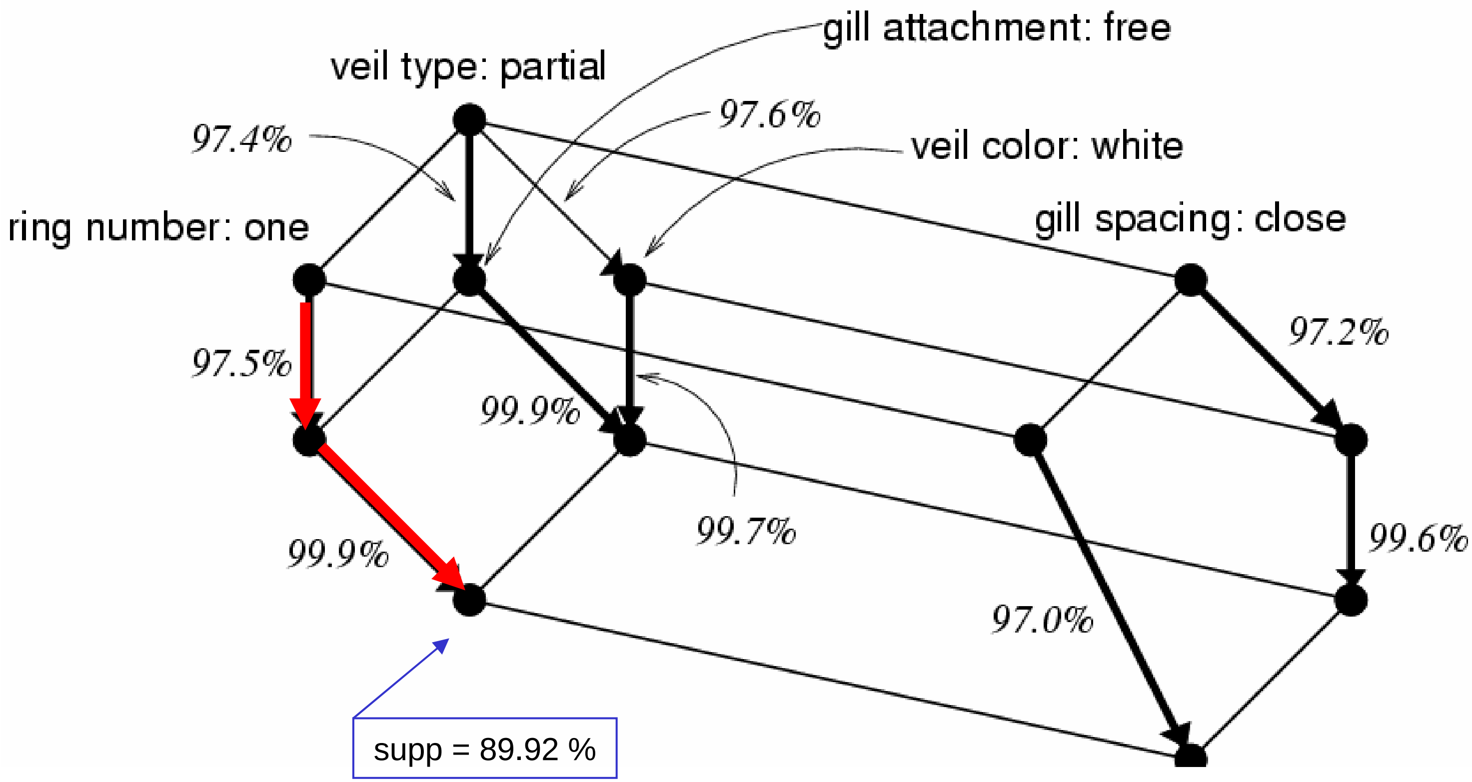
Association rules can be visualized in the iceberg concept lattice:

- **exact rules**
- **approximate rules**

conf = 100 %

conf < 100 %

Luxenburger Basis for approximate association rules



{ring number: one} → {veil color: white}

supp = 89.92 % conf = 97.5 % × 99.9 % ≈ 97.4 %.

Name	Number of objects	Average size of objects	Number of items
T10I4D100K	100,000	10	1,000
MUSHROOMS	8,416	23	127
C20D10K	10,000	20	386
C73D10K	10,000	73	2,177

Some experimental results

Dataset (Minsupp)	Exact rules	D.-G. basis	Minconf	Approximate rules	Luxenburger basis
T10I4D100K (0.5%)	0	0	90%	16,269	3,511
			70%	20,419	4,004
			50%	21,686	4,191
			30%	22,952	4,519
MUSHROOMS (30%)	7,476	69	90%	12,911	563
			70%	37,671	968
			50%	56,703	1,169
			30%	71,412	1,260
C20D10K (50%)	2,277	11	90%	36,012	1,379
			70%	89,601	1,948
			50%	116,791	1,948
			30%	116,791	1,948
C73D10K (90%)	52,035	15	95%	1,606,726	4,052
			90%	2,053,896	4,089
			85%	2,053,936	4,089
			80%	2,053,936	4,089

1. Motivation: Structuring the Frequent Itemset Space

2. Formal Concept Analysis

3. Conceptual Clustering with Iceberg Concept Lattices

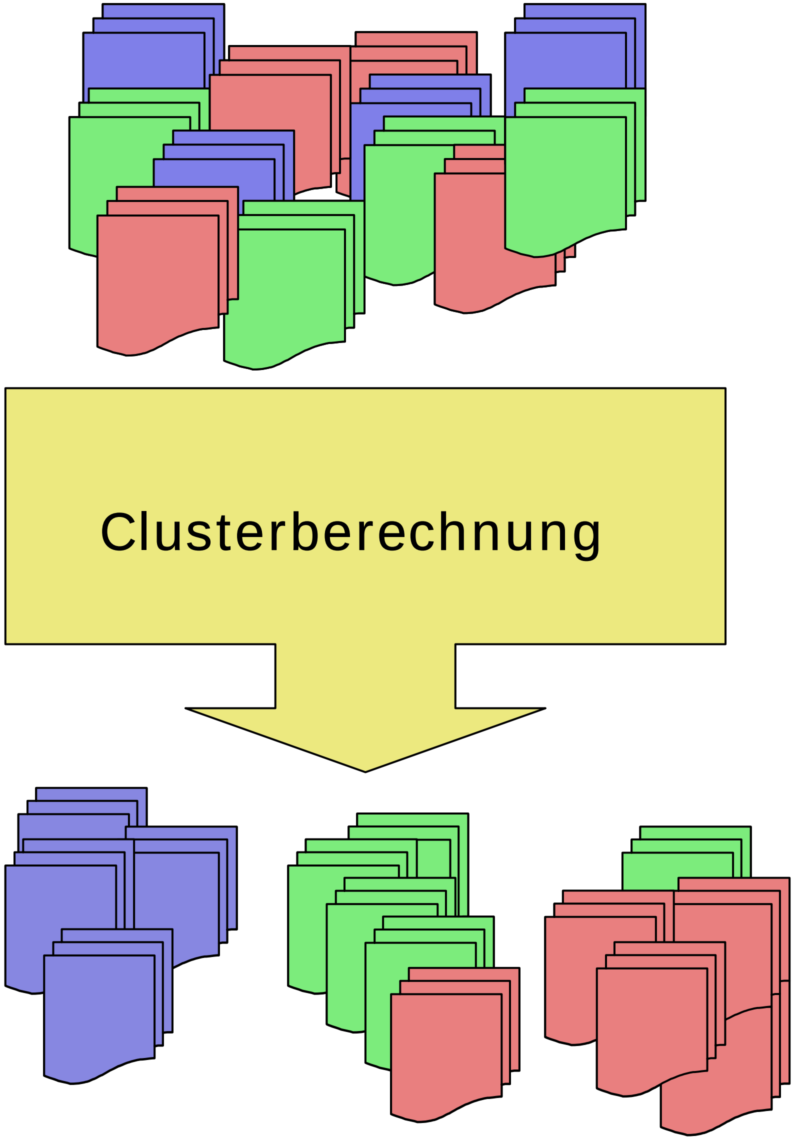
4. FCA-Based Mining of Association Rules

5. **Text Clustering with Background Knowledge**

Joint work with L. Lakhal, Y. Bastide, N. Pasquier, R. Taouil.

Joint work of A. Hotho + G. Stumme

(Begriffliches) Clustern



Dokumente

Cluster (mit Beschreibungen)

Clustern von Texten mit Hintergrundwissen

Aufgabe beim Clustern:

Zusammenfassen von ähnlichen Objekten zu Gruppen (Clustern).

Test-Daten:

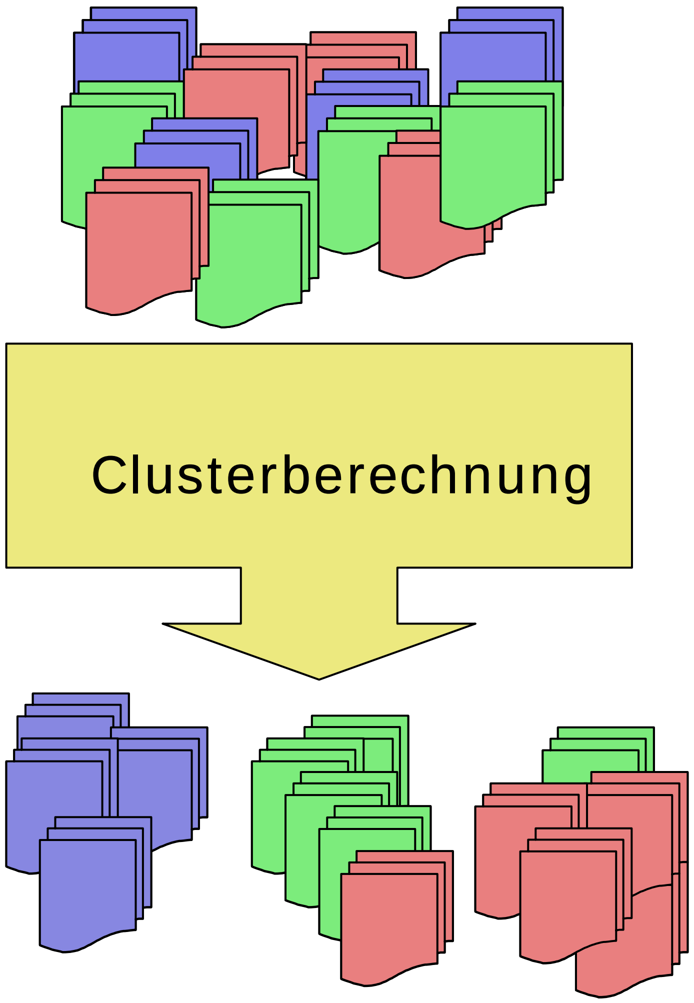
(Eine Teilmenge von) 21578 Reuters-Nachrichtentexten

Problem:

1. Überlappende Cluster sollen erlaubt sein.
2. Beschreibung der Cluster erwünscht.
3. Verfahren soll effizient sein.

Zusatzfrage:

Kann Hintergrundwissen das Ergebnis verbessern?



Verschiedene Ansätze

Formale Begriffsanalyse

+ bietet intensionale Beschreibung
+ Dokumente können zu mehreren Clustern gehören
— Berechnung ist teuer
— evtl. „Overfitting“

Partitionierendes Clustern (z.B. k-Means)

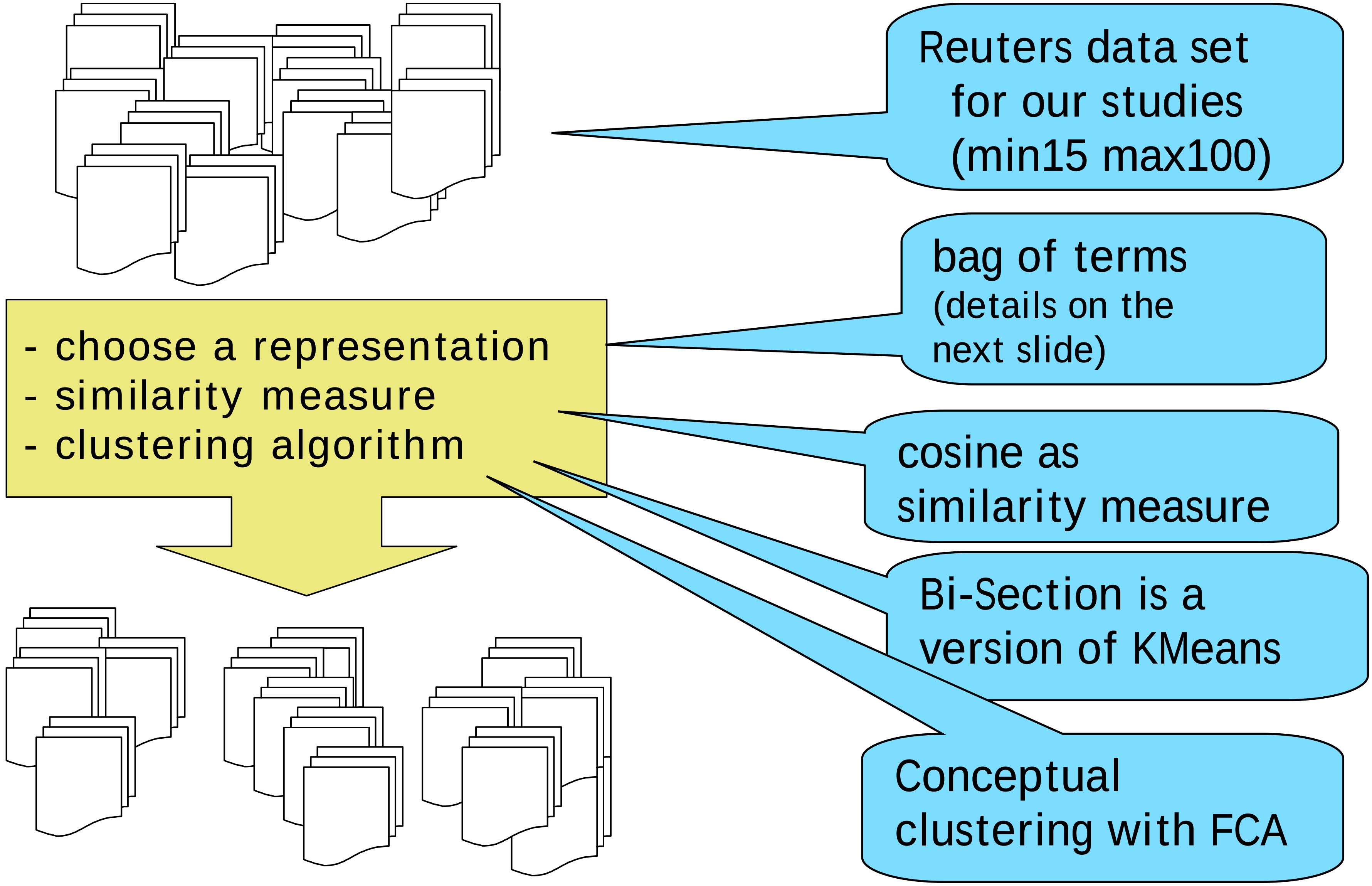
+ clustert große Datenmengen schnell
— die Ergebnisse sind für Menschen schwer verständlich

Begriffliches Clustern

- **Kombination von FBA und Standard Text-Clustering**

- Vorverarbeitung der Dokumente
- Anreicherung mit Hintergrundwissen (Wordnet)
- **Bestimmen einer geeigneten Zahl *k* von Clustern mit *k*-Means**
- Extraktion von Beschreibungen der Cluster
- **Weitere Clusterung mit Begriffsanalyse**
- Visualisierung der Cluster im Begriffsverband

Text Clustering mit Hintergrundwissen



Preprocessing steps

- build a bag of words model

docid	term1	term2	term3	...
doc1	0	0	1	
doc2	2	3	1	
doc3	10	0	0	
doc4	2	23	0	
...				

- extract word counts (term frequencies)
- remove stopwords
- pruning: drop words with less than e.g. 30 occurrences
- weighting of document vectors with tfidf

  (term frequency - inverted document frequency)

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

$|D|$ no. of documents d

$df(t)$ no. of documents d which contain term t

The Bag-of-Words-Model – the Classical Approach

- The bag-of-words-model is the standard feature representation for content-based text mining.
  - Hypothesis: patterns in terminology reflect patterns in conceptualizations.
  - Steps: chunking, stemming, stop words, weighting… go !
  - Good statistical properties.

[Salton 1989]

- Some known deficiencies:
  - collocations (multi word expressions),
  - synonymous terminology,
  - polysemous terminology, and
  - varying degrees of specificity / generalization.

Limitations of the Bag-Of-Words Model

•Thus, algorithms can only detect patterns in *terminology* -- *conceptual patterns* are ignored.

•Specifically, such systems fail to cope with:

1. Multi Word Expressions: **European Union** vs. **Union**,
2. Synonymous Terminology: **Tungsten** vs. **Wolfram**,
3. Polysemous Terminology: **nut**
4. Generalizations: **beef** vs. **pork**

Our Approach

3. Polysemous

- If we enhance the bag-of-words document representation with appropriate ontology concepts, this should improve classification by addressing issues 1-3.

4. Generalizatio

- If we carefully generalize these concepts, this should improve classification even more by addressing issue 4.

Conceptual Document Representation

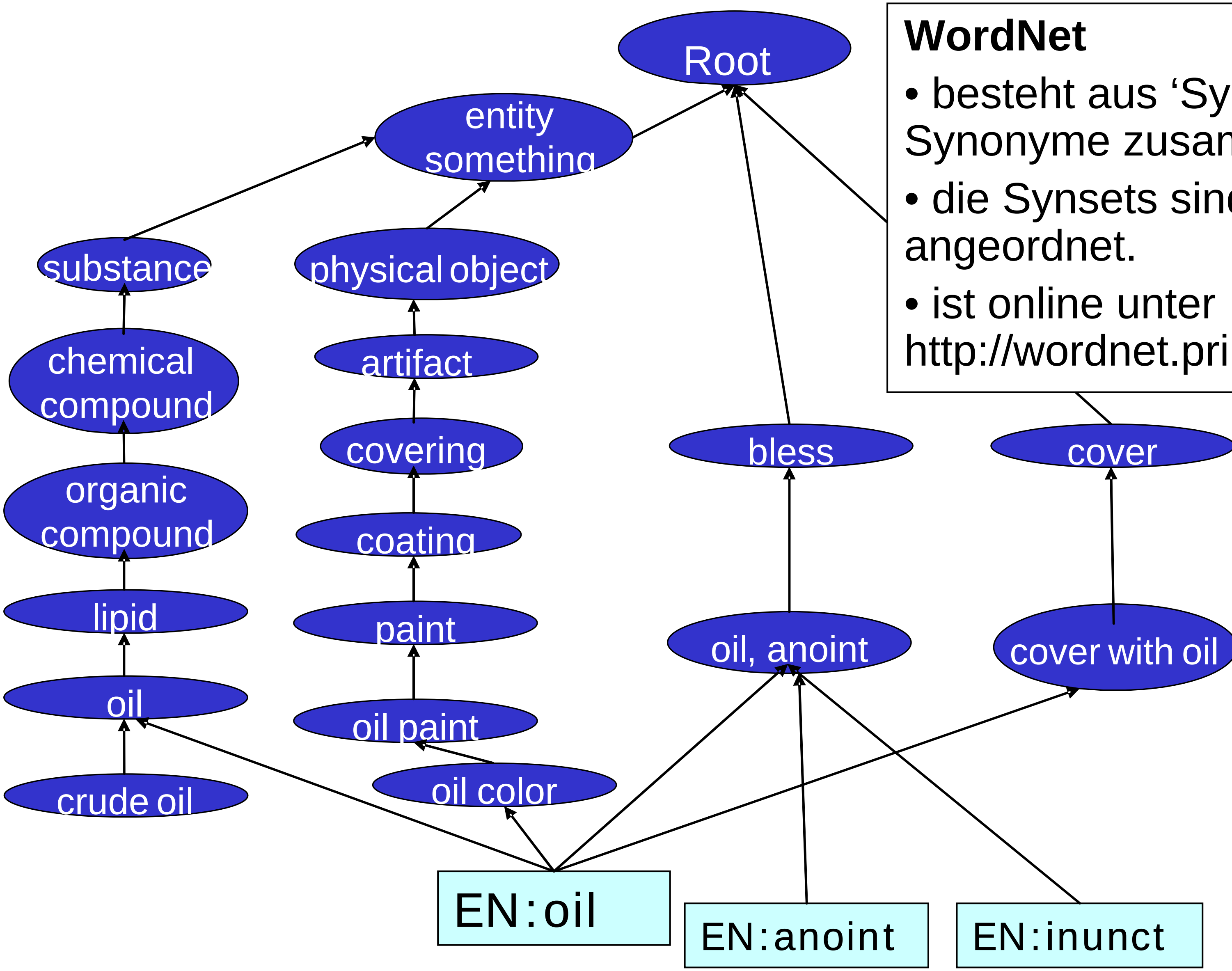
Vorverarbeitung

Test-Daten: Reuters-21578 Corpus

- 1015 Documente ausgewählt, so dass jede Klasse min. 25 und max. 30 Dokumente enthält

- **Vorverarbeitung**
- “Bag of words” Modell
- Stopworte entfernen
- Seltene Worte (<5) entfernen
- Hinzufügen genereller Terme mit WordNet

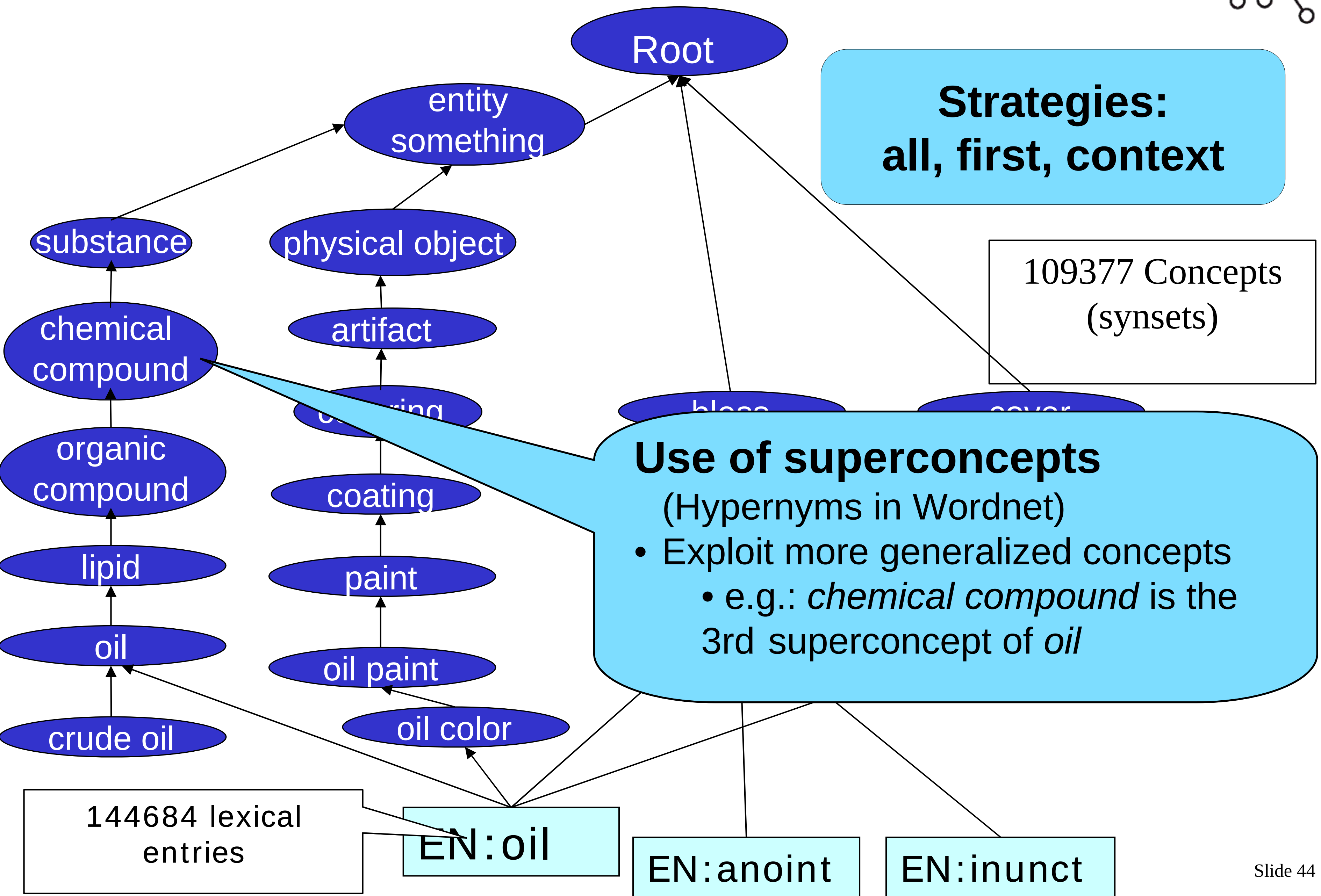
WordNet



WordNet

- besteht aus ‘Synsets’, die Synonyme zusammenfassen.
- die Synsets sind hierarchisch angeordnet.
- ist online unter http://wordnet.princeton.edu

Hinzufügen von Oberbegriffen aus WordNet



Strategies:
all, first, context

Use of superconcepts
(Hypernyms in Wordnet)
- Exploit more generalized concepts
  - e.g.: *chemical compound* is the 3rd superconcept of *oil*

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
...	...

Clustern von Texten mit Hintergrundwissen

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
...	...
chem. comp.	2

Hinzufügen von Oberbegriffen aus WordNet

Clustern von Texten mit Hintergrundwissen

- **Zweistufiger Cluster-Ansatz:**
  - Erster Cluster-Schritt:
    - mit Standard-Algorithmus “Bisection k-Means”
    - reduziert effizient die Anzahl der Objekte
  - Zweiter Cluster-Schritt:
    - mit Formaler Begriffsanalyse
    - liefert intensionale Beschreibungen der Cluster
    - und erlaubt Mehrfachvererbung

1. Schritt: Partitionierendes Clustern

Partitionierender Cluster-Algorithmus

- Bi-Section Version von *k*-Means
- Kosinus als Ähnlichkeitsmaß

Bi-Partitioning K-Means

- Input: Set of documents *D*, number of clusters *k*
- Output: *k* cluster that exhaustively partition *D*

- Initialize: *P* = {D}*

- **Outer Loop**:
  Repeat *k-1* times: **Bi-Partition** the largest cluster $E \in P^*$

-

Bi-Partitioning K-Means

- Input: Set of documents D, number of clusters k
- Output: k cluster that exhaustively partition D

- Initialize: $P^* = \{D\}$

- **Outer loop**:
  Repeat k-1 times: **Bi-Partition** the largest cluster $E \in P^*$

- **Inner loop**:
  - Randomly initialize two documents from E to become e_1, e_2
  - **Repeat** until convergence is reached
    - **Assign each** document **from *E* to the** nearest **of the two** e_i **; thus split E into** $\boldsymbol{E_1, E_2}$
    - Re-compute e_1, e_2 **to become the centroids of the document representations assigned to them**
  - $P^* := (P^* \setminus E) \cup \{E_1, E_2\}$

2. Schritt Begriffliches Clustern

Partitionierender Cluster-Algorithmus

- wie oben beschrieben

- **Extraktion von Cluster-Beschreibungen**
- die Verwendung aller Synsets erzeugt einen zu großen Verband
- Auswahl jeweils der Synsets, die für das Cluster über einem gegebenen Schwellwert θ liegen

Begriffliches Clustern mit Begriffsanalyse

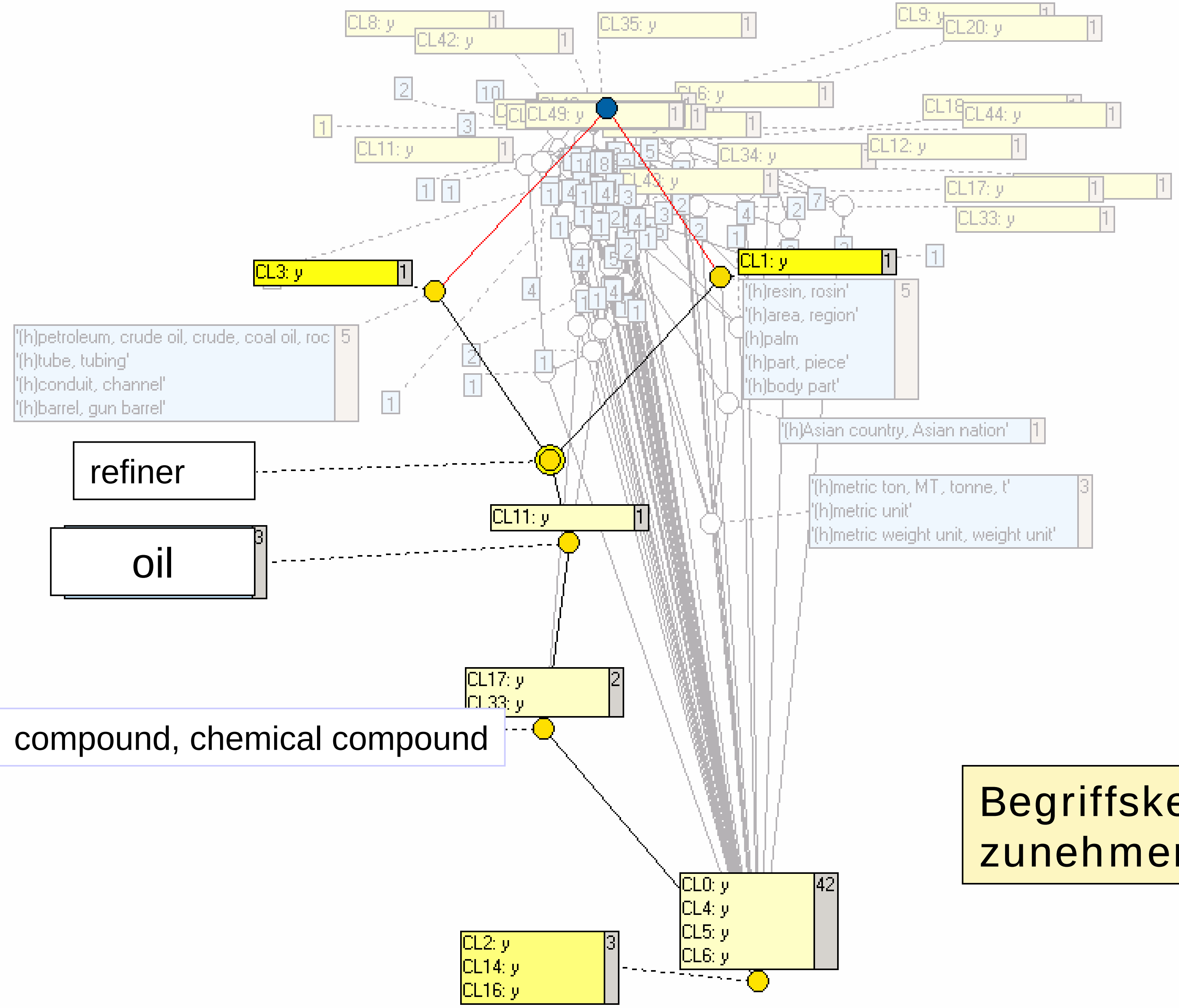
- Berechnung des Begriffsverbandes erzeugt intensionale Beschreibungen der Cluster
- Visualisierung

Extracted Word description

Cluster 0		Cluster 1		Cluster 2		Cluster 3		Cluster 4	
amount	0,12	depository financial institu	0,09	loss	0,34	Irani, Iranian, Persian'	0,14	indebtedness, liability, fin	0,12
billion, one million million,	0,11	financial institution, financ	0,09	failure	0,33	Iran, Islamic Republic of	0,13	obligation	0,12
large integer'	0,11	rate, charge per unit'	0,09	nonaccomplishment, nona	0,32	gulf	0,13	debt	0,12
integer, whole number'	0,11	charge	0,09	Connecticut, Nutmeg Sta	0,28	vessel, watercraft'	0,12	written agreement'	0,1
insufficiency, inadequacy	0,1	institution, establishment'	0,09	ten, 10, X, tenner, decade	0,24	ship	0,12	agreement, understanding	0,08
deficit, shortage, shortfall	0,1	loss	0,08	American state'	0,23	craft	0,12	creditor	0,08
number	0,09	monetary unit'	0,07	state, province'	0,22	Asian, Asiatic'	0,11	lender, loaner'	0,08
excess, surplus, surplusa	0,09	central, telephone exchan	0,07	system, unit'	0,19	person of color, person of	0,10	statement	0,07
overabundance, overmuc	0,09	financial loss'	0,06	network, net, mesh, mesh	0,19	Asian country, Asian natio	0,10	billion, one million million,	0,06
abundance, copiousness,	0,09	outgo, expenditure, outlay	0,06	September, Sep, Sept'	0,18	oil tanker, oiler, tanker, ta	0,10	large integer'	0,05

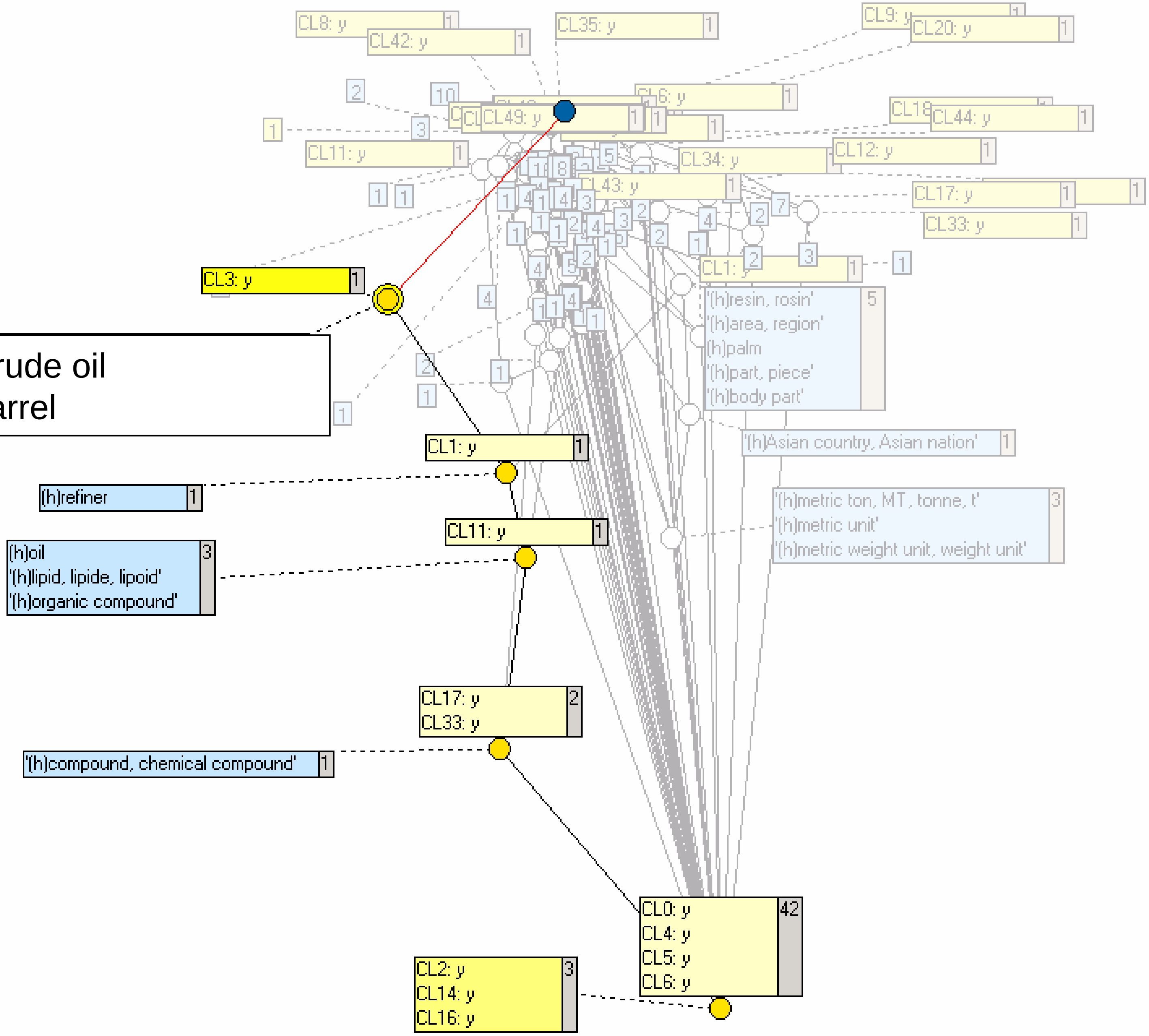
Cluster 5		Cluster 6		Cluster 7		Cluster 8		Cluster 9	
text, textual matter'	0,15	loss	0,34	gross sales, gross revenu	0,11	tender, legal tender'	0,15	metric weight unit, weight	0,15
matter	0,15	failure	0,33	sum, sum of money, amo	0,09	offer, offering'	0,14	metric ton, MT, tonne, t'	0,15
letter, missive'	0,15	nonaccomplishment, nona	0,32	income	0,09	medium of exchange, mon	0,11	mass unit'	0,14
sign, mark'	0,13	common fraction, simple f	0,22	financial gain'	0,09	speech act'	0,1	palm, thenar'	0,14
clue, clew, cue'	0,13	fraction	0,22	gain	0,09	indicator	0,1	area, region'	0,12
purpose, intent, intention	0,11	rational number'	0,22	enterprise	0,05	standard, criterion, measu	0,1	unit of measurement, unit	0,10
evidence	0,11	real number, real'	0,22	business, concern, busine	0,05	reference point, point of r	0,09	organic compound'	0,10
indication, indicant'	0,11	complex number, complex	0,22	assets	0,05	signal, signaling, sign'	0,08	oil	0,10
goal, end'	0,1	one-half, half'	0,22	division	0,05	acquisition	0,06	lipid, lipide, lipoid'	0,10
writing, written material, p	0,07	revolutions per minute, rp	0,22	army unit'	0,05	giant	0,06	compound, chemical com	0,08

Ergebnisse



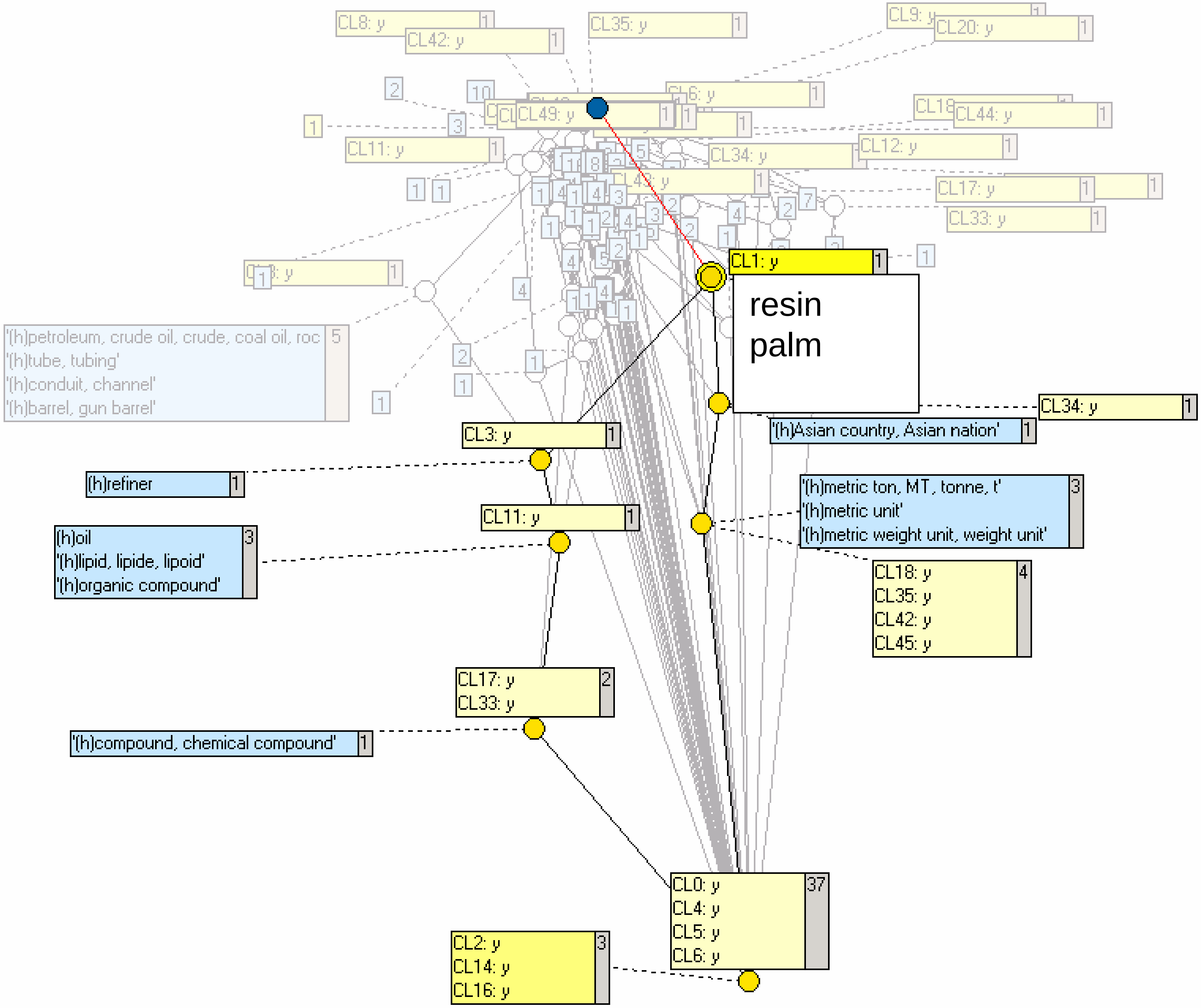
Begriffskette mit zunehmender Spezifität

Ergebnisse



Crude oil
barrel

Ergebnisse


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

Literatur

- Stephan Bloehdorn, Andreas Hotho: *Text Classification by Boosting Weak Learners based on Terms and Concepts*. ICDM 2004.
- Andreas Hotho, Steffen Staab, Gerd Stumme: *WordNet improves text document clustering*; Semantic Web Workshop @ SIGIR 2003.
- Alexander Maedche, Steffen Staab. *Ontology Learning for the Semantic Web*. IEEE Intelligent Systems, 16(2):72–79, 2001.
- Philipp Cimiano, Andreas Hotho, Steffen Staab. *Comparing Conceptual, Partitional and Agglomerative Clustering for Learning Taxonomies from Text*. ECAI 2004. Extended Version to appear (JARS 2005).

1. **Motivation: Structuring the Frequent Itemset Space**
2. **Formal Concept Analysis**
3. **Conceptual Clustering with Iceberg Concept Lattices**
4. **FCA-Based Mining of Association Rules**
5. **Text Clustering with Background Knowledge**