VII.5 Formale Begriffsanalyse

Formale Begriffsanalyse ist um 1980 als mathematische Theorie entstanden, die eine Formalisierung des Begriffs vom „Begriff" liefert.

FBA hat seitdem zunehmend Verbreitung in der Informatik gefunden, u.a. in

- der Datenanalyse,
- der Wissensentdeckung,
- dem Software Engineering.

Ausgehend von Datensätzen leitet FBA Begriffshierarchien ab.

FBA ermöglicht die Erzeugung und die Visualisierung der Begriffshierarchien.

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.	Kaffee-qualität	Technische Prüfung	Sicherheit	Handhabung	test-Qualitätsurteil
Gewichtung			35 %	30 %	10 %	25 %	
Neckermann Best.-Nr. 8628/409	40,–	35,–*) / ❑	baugl. mit Otto Hanseatic Best.-Nr. 4327357				zufriedenst.
Otto Hanseatic Best.-Nr. 4327357	40,–	30,–*) / ❑	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

Formal Concept Analysis [Wille 1982]

- 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.
- FCA is used for data analysis, information retrieval, and knowledge discovery.
- FCA can be understood as conceptual clustering method, which clusters simultanously objects and their descriptions.
- FCA can be used for efficiently computing association rules.

VII.5.2 Basics of Formal Concept Analysis

- In its basic version, FCA handles object-attribute pairs.
- Object-attribute-value triples can also be handled, but this is not topic of this course.
- **Def.:** A (**formal**) **context** is a triple (G,M,I) where G and M are sets and I is a binary relation between G and M.
  - The elements of G are called objects, and the elements of M are called attributes.
  - $(g,m) \in I$ is read „object g has attribute m".

Entspricht Items/Transaktionen bei Assoziationsregeln

Def.: Ein **formaler Kontext** ist ein Tripel (G,M,I), wobei

- G eine Menge von Gegenständen,
- M eine Menge von Merkmalen
- und I eine Relation zwischen G und M ist.

- $(g,m) \in I$ wird gelesen als „Gegenstand g hat Merkmal 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	×	×	×	×	×	×	×	×

Für $A \subseteq G$ definieren wir

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

Für $B \subseteq M$ definieren wir dual

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

A´

A

National Parks in California				Swimming	Boating		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	×	×	×	×	×	×		
								×
				×	×			
Yosemite Natl. Park	×	×	×	×	×	×	×	×

Für $A \subseteq G$ definieren wir

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

Für $B \subseteq M$ definieren wir dual

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

B

B‘

National Parks in California				Swimming	Boating	Fishing	Bicycle Trail	Cross Country Trail
Cabrillo Natl. Mon.						×	×	
Channel Islands Natl. Park		×		×		×		
				×			×	
				×		×		
Fort Point Natl. Historic Site	×					×		
				×		×	×	
John Muir Natl. Historic Site	×							
						×		×
				×	×	×		×
Lava Beds Natl. Mon.	×	×						
Muir Woods Natl. Mon.		×						
Pinnacles Natl. Mon.		×						
				×		×	×	
				×		×		
				×	×	×		
						×		×
				×	×	×		
				×	×	×	×	×

Für $A, A_1, A_2 \subseteq G$ gilt:

- $A_1 \subseteq A_2 \Rightarrow A'_2 \subseteq A'_1$
- $A \subseteq A''$
- $A' = A'''$

Für $B, B_1, B_2 \subseteq M$ gilt:

- $B_1 \subseteq B_2 \Rightarrow B'_2 \subseteq B'_1$
- $B \subseteq B''$
- $B' = B'''$

A´

A

National Parks in California				Swimming	Boating		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	×	×	×	×	×	×		
								×
				×	×			
Yosemite Natl. Park	×	×	×	×	×	×	×	×

VII.5 Formale Begriffsanalyse

Def.: Ein **formaler Begriff** ist ein Paar (A,B) mit

- $A \subseteq G$ und $B \subseteq M$,
- $A' = B$,
- $B' = A$.

A ist der **Umfang** und B der **Inhalt** des Begriffs.

Inhalt

Umfang

National Parks in California				Swimming	Boating		Bicycle Trail	Cross Country Trail
Cabrillo Natl. Mon.						×	×	
Channel Islands Natl. Park		×		×		×		
Death Valley Natl. Mon.	×	×	×	×			×	
				×				
Fort Point Natl. Historic Site	×					×		
				×		×	×	
John Muir Natl. Historic Site	×							
Joshua Tree Natl. Mon.	×	×	×					
								×
				×	×			×
Lava Beds Natl. Mon.	×	×						
Muir Woods Natl. Mon.		×						
Pinnacles Natl. Mon.		×						
				×			×	
				×				
				×	×			
								×
				×	×			
				×	×		×	×

VII.5 Formale Begriffsanalyse

Der blaue Begriff ist ein **Unterbegriff** des gelben Begriffs, denn

der blaue Umfang ist im gelben Umfang enthalten.

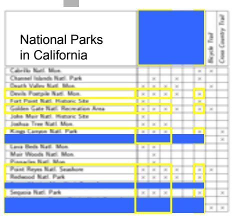
(⇔ der gelbe Inhalt ist im blauen Inhalt enthalten.)

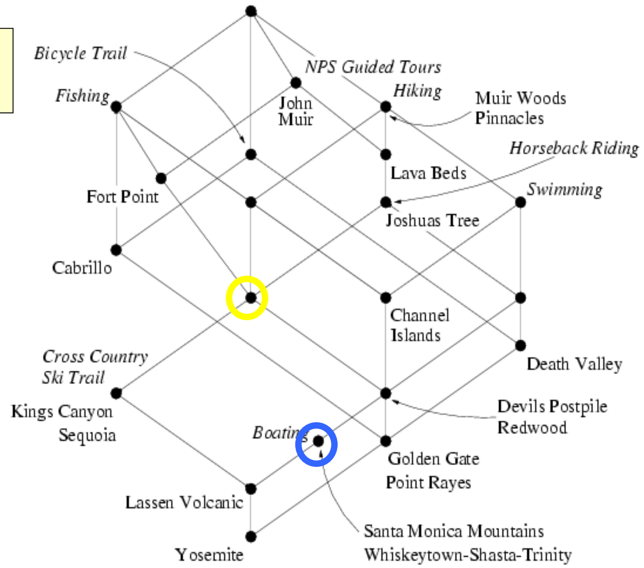
$(A,B) \leq (C,D) :\Leftrightarrow A \subseteq C$
$(\Leftrightarrow B \supseteq D)$

National Parks in California							Bicycle Trail	Cross Country Trail
Cabrillo Natl. Mon.						×	×	
Channel Islands Natl. Park		×		×		×		
Death Valley Natl. Mon.	×	×	×	×			×	
				×				
Fort Point Natl. Historic Site	×					×		
				×			×	
John Muir Natl. Historic Site	×							
Joshua Tree Natl. Mon.	×	×	×					
								×
								×
Lava Beds Natl. Mon.	×	×						
Muir Woods Natl. Mon.		×						
Pinnacles Natl. Mon.		×						
				×			×	
				×				
								×
							×	×

VII.5 Formale Begriffsanalyse

Der **Begriffsverband** zu dem Nationalpark-Kontext





VII.5 Formale Begriffsanalyse

- **Def.:** The **concept lattice** [**Begriffsverband**] of a formal context (G,M,I) is the set of all formal concepts of (G,M,I), together with the partial order

  $(A_1,B_1) \leq (A_2,B_2) :\Leftrightarrow A_1 \subseteq A_2 \quad (\Leftrightarrow B_1 \supseteq B_2)$.

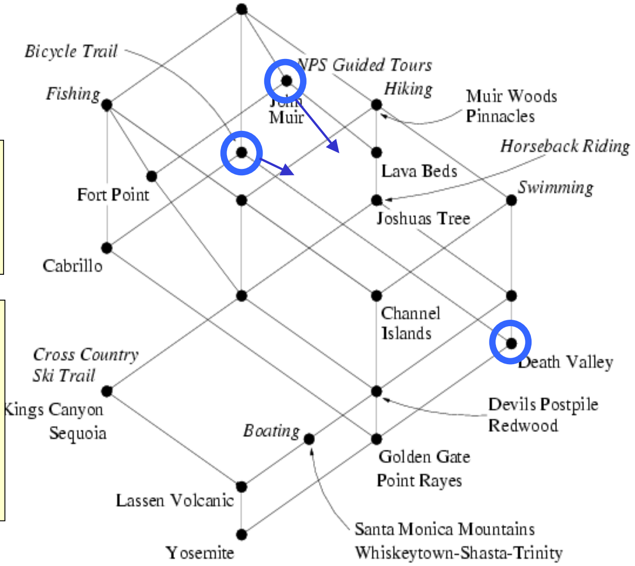
  The concept lattice is denoted by $\underline{\mathfrak{B}}(G,M,I)$.

- **Theorem:** The concept lattice is a lattice, i.e. for two concepts (A_1,B_1) and (A_2,B_2), there is always
  - a greatest common subconcept: $(A_1 \cap A_2 , (B_1 \cup B_2)'')$
  - and a least common superconcept: $((A_1 \cup A_2)'', B_1 \cap B_2)$.

„Welche Gegenstände haben sowohl das Merkmal ‚Bicycle Trail' als auch ‚NPS Guided Tours'?"

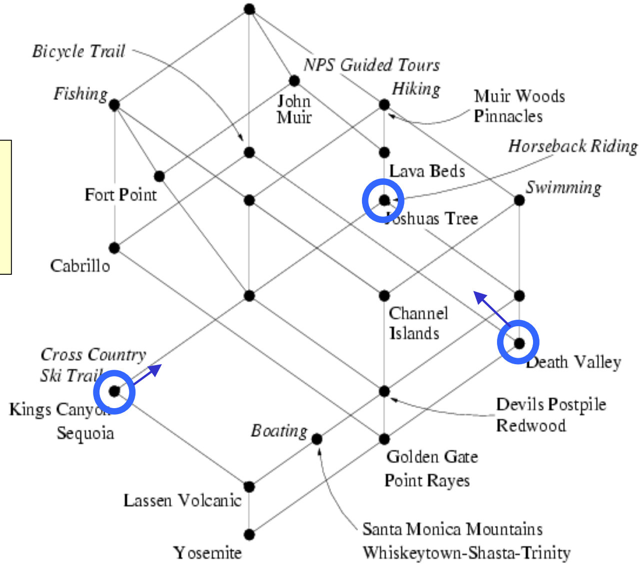
„Welche Merkmale haben diese Gegenstände noch?"

In anderen Worten: „Welche Merkmale folgen noch aus ‚Bicycle Trail' und ‚NPS Guided Tours'?"



„Welche Merkmale teilen sich die Gegenstände ‚Kings Canyon' und ‚Death Valley'?"

„Welche Gegenstände haben noch diese Merkmale?"



VII.5.3 Formal Concept Analysis as Conceptual Clustering Method

- **Conceptual Clustering** methods are clustering methods which generate simultaneously descriptions of the clusters.

- Advantages of conceptual clustering against clustering as in Sect. VII.1:
  - A cluster is not only a set of objects, but there also exists an intensional description.
  - For FCA: The results do not depend on the order of the input

- Disadvantages:
  - The language used to describe the clusters restricts the type of clusters which can be built.
  - The computation has usually higher complexity.

- Other methods: Michalski & Stepp 1983; Lebowitz 1987; Fisher 1987; Gennari et al 1989

Iceberg concept lattices only allow conjunctions of attributes as descriptions.

- In the notion of Formal Concept Analysis, the support of an itemset $X \subseteq M$ can be written as

$$\text{supp}(X) = \frac{|X'|}{|G|}$$

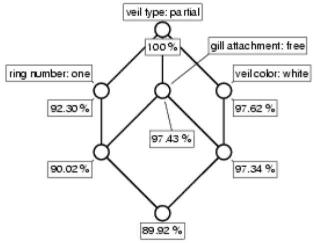
- Def.: The **iceberg concept lattice** of a formal context (G,M,I) for a given minimal support minsupp is the set

$$\{(A,B) \in \underline{\mathbf{B}}(G,M,I) \mid \text{supp}(B) \geq \text{minsupp}\}$$

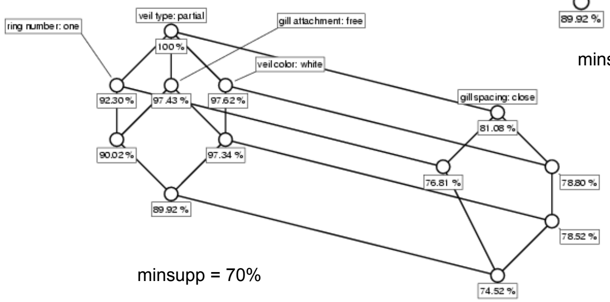
- It can be computed with the algorithm **TITANIC**. [Stumme et al 2001]

Example:

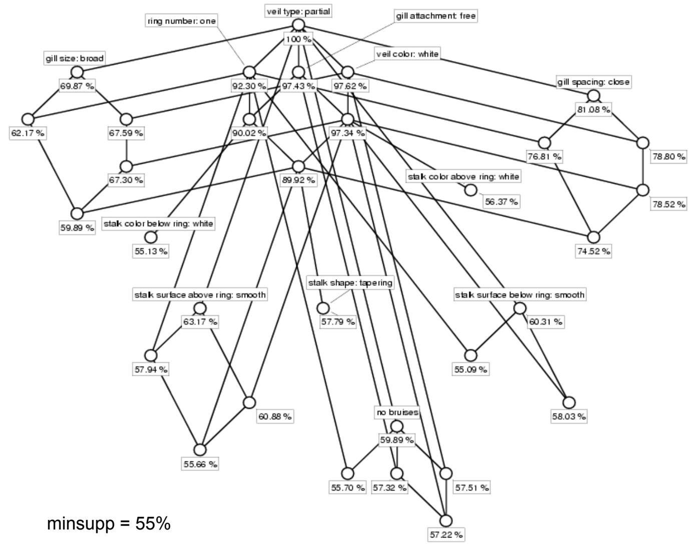
Iceberg Concept Lattices of the Mushrooms Database (http:\\kdd.ics.uci.edu).



minsupp = 85%



minsupp = 70%



minsupp = 55%

VII.5.4 Computation of (Iceberg) Concept Lattices

- There exist a number of algorithms for computing concept lattices
  - Next-Closure [Ganter 1984]
  - Titanic [Stumme et al 2001]
  - and some incremental algorithms
- The following method is also suitable for manual computation. [Wille 1982]
- It provides the best worst-case time complexity. [Nourine, Raynoud 1999]

↳ Example „Faces“ on the Blackboard

How to compute/draw a concept lattice:

- From left to right, consider all intersections of each column extent with every column extent to the left of it. If the resulting extent is not already a column, add it as column at the right end of the context. Repeat this until the last (added) column is reached.
- Add a full column, unless there is already one. (Now each column stands for one concept.)
- Draw a circle for the full column.
- Draw for each column, starting for the ones with a maximal number of crosses, a circle, and link it with a line to the circles where the column comprises the current column.
- Attach every attribute label to the circle of the corresponding column.
- Attach every object label to the circle laying exactly below the circles of the attributes in its intent.

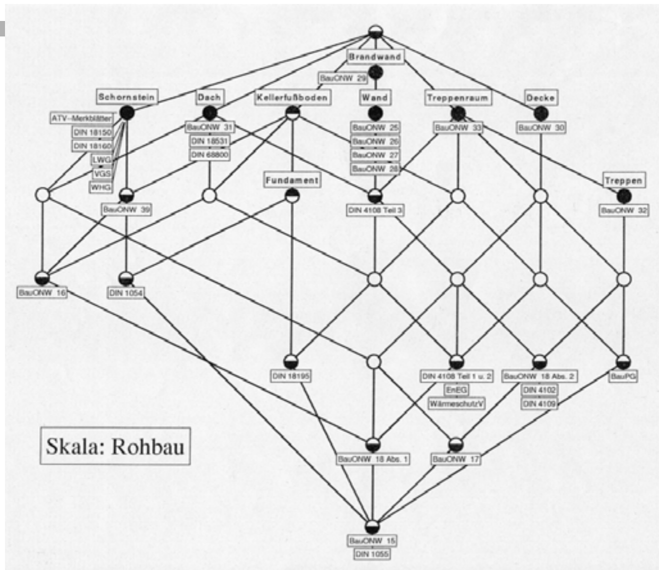
How to check the drawing of a concept lattice:

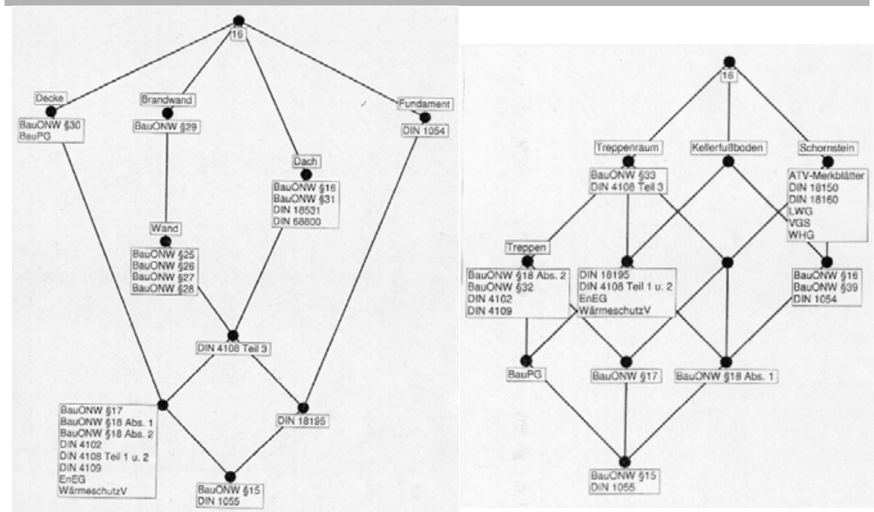
- Is it really a lattice? (This test is usually skipped.)
- Is every concept with exactly one upper neighbor labeled by at least one attribute?
- Is every concept with exactly one lower neighbor labeled by at least one object?
- Is, for all $g \in G$ and all $m \in M$, the label of object g below the label of attribute m iff $(g,m) \in I$?

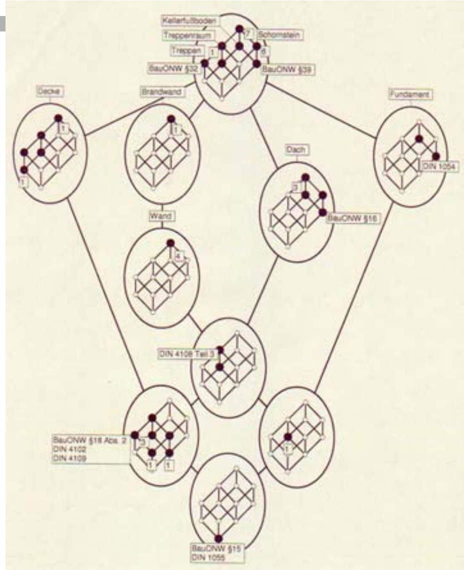
- Another method for reducing the complexity of the diagram is conceptual scaling.
- In our case (only Boolean attributes), this means selecting only subsets of the attribute set.
- The resulting concept lattices are smaller.
- If combinations are of interest, they can be put together again.
- Conceptual scaling can be comgined with Iceberg concept lattices.

Beispiel: Baurecht in Nordrhein-Westfalen

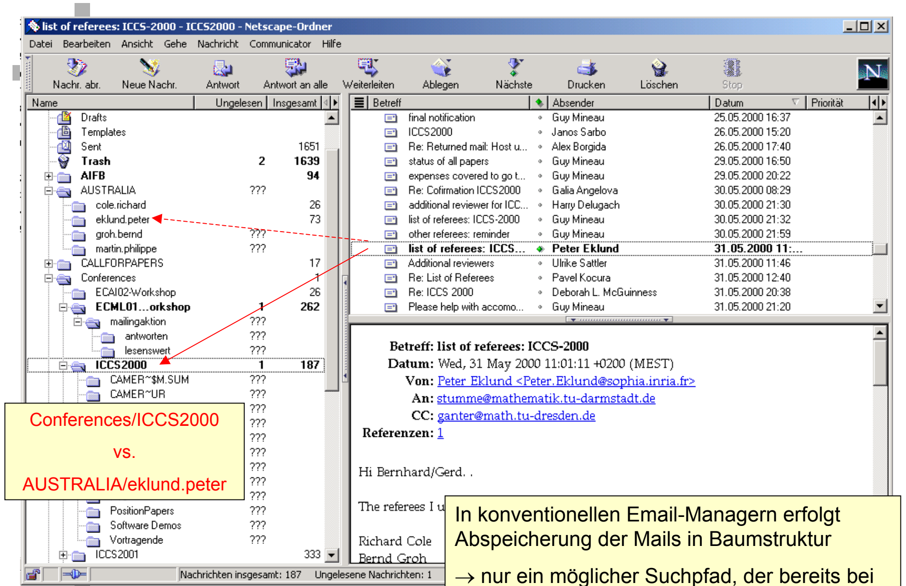
	Dach	Decke	Wand	Brandwand	Treppen	Treppenraum	Fundament	Kellerfußboden	Schornstein
BauONW 15	X	X	X	X	X	X	X	X	X
BauONW 16	X							X	X
BauONW 17	X	X	X	X	X	X			X
BauONW 18 Abs. 1	X	X	X	X		X		X	X
BauONW 18 Abs. 2	X	X	X	X	X	X			
BauONW 25			X	X					
BauONW 26			X	X					
BauONW 27			X	X					
BauONW 28			X	X					
BauONW 29				X					
BauONW 30		X							
BauONW 31	X								
BauONW 32					X	X			
BauONW 33						X			
BauONW 36									
BauONW 39								X	X
BauONW 40									
BImSchG									
BauPG		X			X	X		X	
EnEG	X	X	X	X		X		X	
WHG									X
LWG									X
WärmeschutzV	X	X	X	X		X		X	
HeizAnlV									
BImSchV									
VGS									X
DIN 1054							X	X	X
DIN 1055	X	X	X	X	X	X	X	X	X
DIN 4102	X	X	X	X	X	X			
DIN 4108 Teil 1 u. 2	X	X	X	X		X		X	
DIN 4108 Teil 3	X		X	X		X			
DIN 4109	X	X	X	X	X	X			
DIN 18150									X
DIN 18160									X
DIN 18195	X		X	X		X	X	X	
DIN 18531	X								
DIN 68800	X								
DIN-Normen für Feuerungsanlagen									
DIN-Normen für Entwässerung									
ATV-Merkblätter									X







- Some **typical applications** of FCA:
  - analysis of children suffering from diabetes
  - IT security management system
  - database marketing in a Suiss department store
  - Conceptual Email Management system
  - developing qualitative theories in music estethics
  - analysis of flight movements at Frankfurt airport



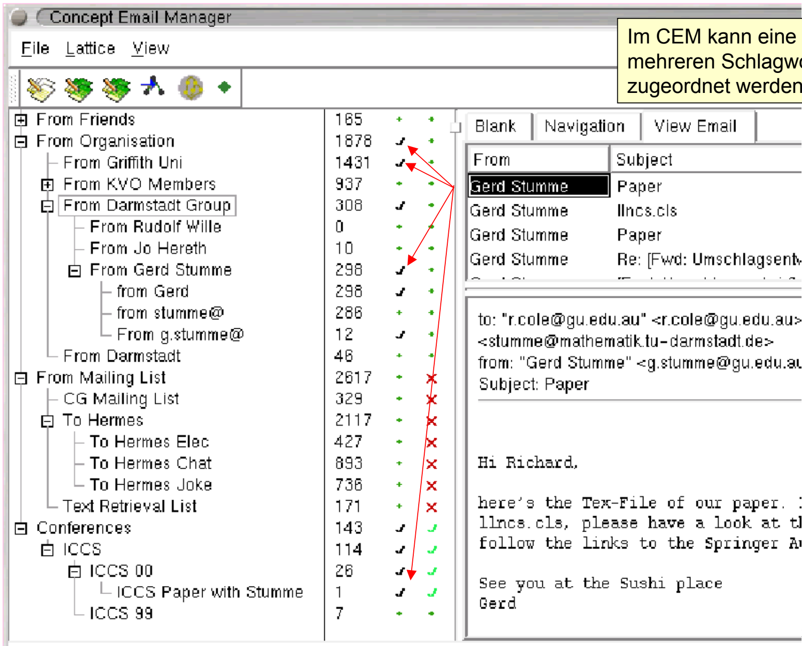
Conferences/ICCS2000

vs.

AUSTRALIA/eklund.peter

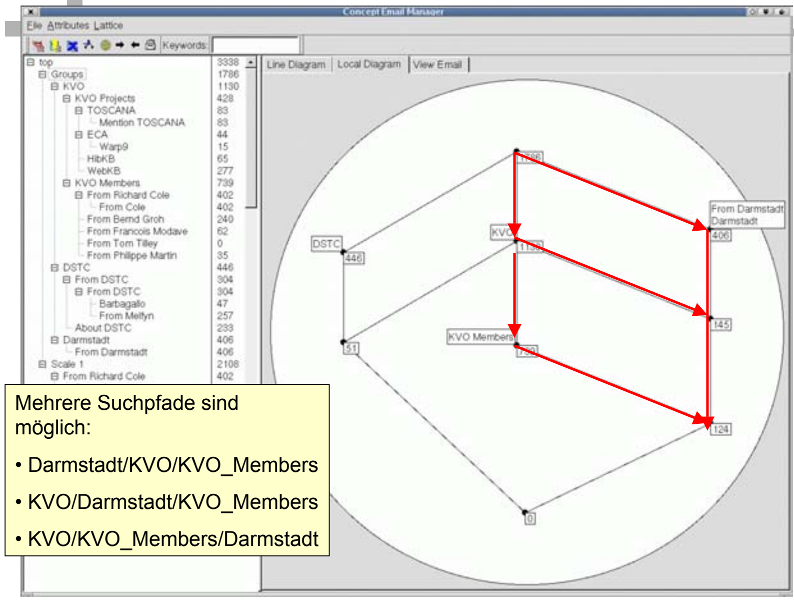
In konventionellen Email-Managern erfolgt Abspeicherung der Mails in Baumstruktur

→ nur ein möglicher Suchpfad, der bereits bei Abspeicherung festgelegt werden muss



Im CEM kann eine Email mehreren Schlagworten zugeordnet werden.

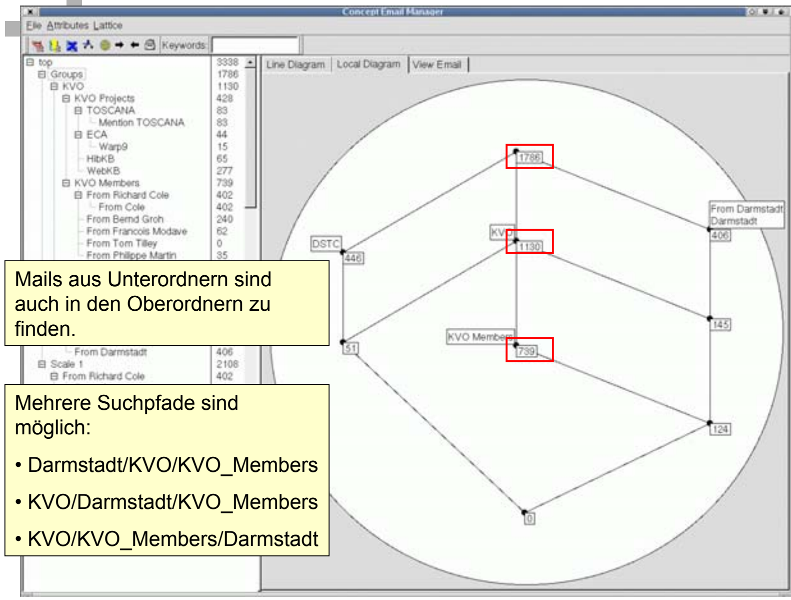
Browsing basierend auf Formaler Begriffsanalyse



Mehrere Suchpfade sind möglich:

- Darmstadt/KVO/KVO_Members
- KVO/Darmstadt/KVO_Members
- KVO/KVO_Members/Darmstadt

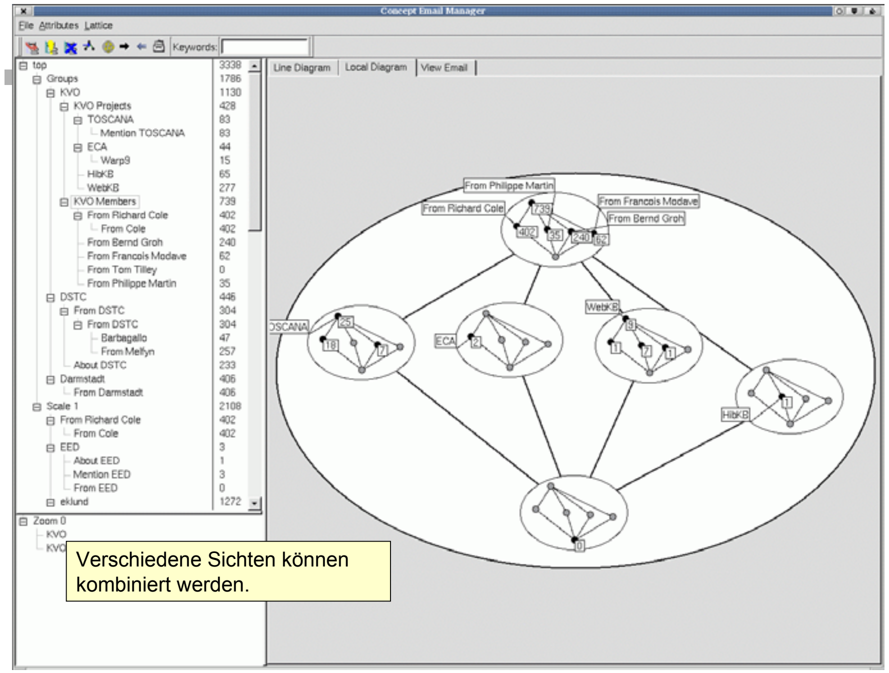
Browsing basierend auf Formaler Begriffsanalyse



Mails aus Unterordnern sind auch in den Oberordnern zu finden.

Mehrere Suchpfade sind möglich:

- Darmstadt/KVO/KVO_Members
- KVO/Darmstadt/KVO_Members
- KVO/KVO_Members/Darmstadt



Verschiedene Sichten können kombiniert werden.

VII.5.5 A reduced representation of association rules based on Formal Concept Analysis

- The input data of association rules algorithms can be written as a formal context (G,M,I): M is the set of items, G consists of the transaction IDs, and the relation I is the list of transactions.

- We will distinguish between exact and approximate association rules:

- **Def.:** An **association rule** $X \rightarrow Y$ (with $X,Y \subseteq M$) is called **exact** if conf($X \rightarrow Y$) = 1 and **approximate** else.

An exact association rule is also called an **implication**.

- In concept lattices, **exact association rules** can be directly read from the diagram:

- **Lemma:** An implication $X \rightarrow Y$ holds iff the largest concept which is below all concepts generated by the attributes in X is below all concepts generated by attributes in Y.

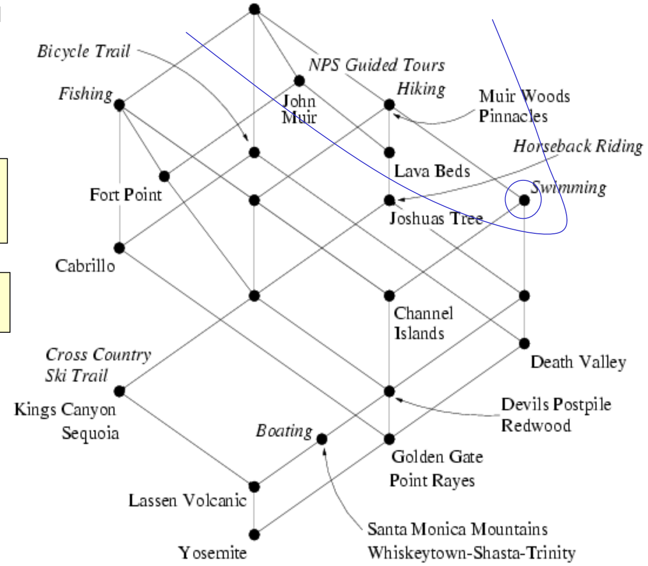
- **Examples:**
  - {Swimming} $\rightarrow$ {Hiking}

    (supp=10/19 $\approx$ 52.6%, conf = 100%)

  -

„Welche Merkmale folgen noch aus ‚Swimming'?"

„Hiking, (NPS Guided Tours)"



- In concept lattices, **exact association rules** can be directly read from the diagram:

- **Lemma:** An implication $X \rightarrow Y$ holds iff the largest concept which is below all concepts generated by the attributes in X is below all concepts generated by attributes in Y.

- **Examples:**
  - {Swimming} $\rightarrow$ {Hiking}

    (supp=10/19 $\approx$ 52.6%, conf = 100%)

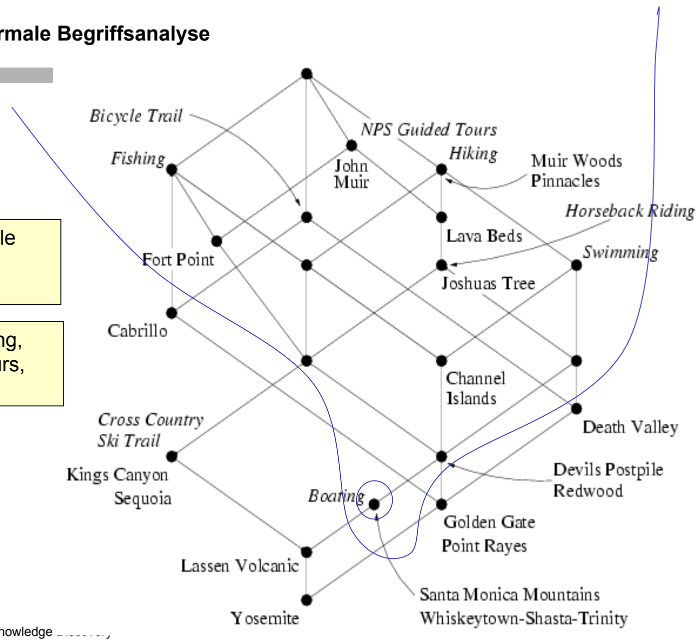
  - {Boating} $\rightarrow$ {Swimming, Hiking, NPS Guided Tours, Fishing}

    (supp=4/19 $\approx$ 21.0%, conf = 100%)

  -

„Welche Merkmale folgen noch aus ‚Boating'?"

„Swimming, Hiking, NPS Guided Tours, Fishing"



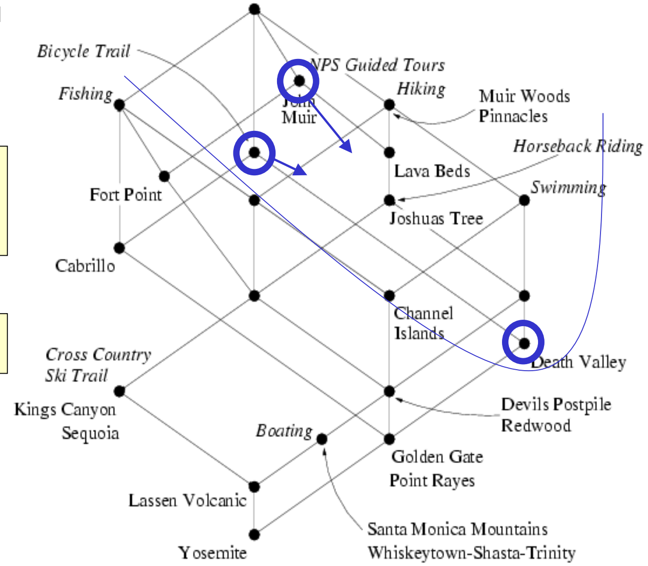
- In concept lattices, **exact association rules** can be directly read from the diagram:

- **Lemma:** An implication $X \rightarrow Y$ holds iff the largest concept which is below all concepts generated by the attributes in X is below all concepts generated by attributes in Y.

- **Examples:**
  - {Swimming} $\rightarrow$ {Hiking}
    (supp=10/19 $\approx$ 52.6%, conf = 100%)
  - {Boating} $\rightarrow$ {Swimming, Hiking, NPS Guided Tours, Fishing}
    (supp=4/19 $\approx$ 21.0%, conf = 100%)
  - {Bicycle Trail, NPS Guided Tours} $\rightarrow$ {Swimming, Hiking}
    (supp=4/19 $\approx$ 21.0%, conf = 100%)

„Welche Merkmale folgen noch aus ‚Bicycle Trail' und ‚NPS Guided Tours'?"

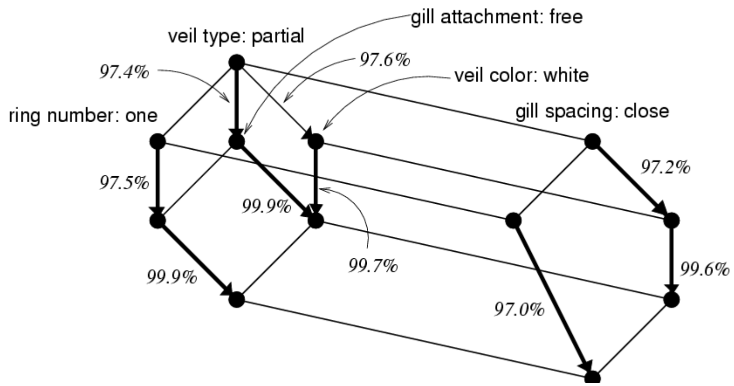
„Swimming, Hiking, Horseback Riding"



- **Lemma:** For $A \subseteq G$ *and* $B \subseteq M$, the following holds:
  - $A_1 \subseteq A_2 \Rightarrow A'_2 \subseteq A'_1$
  - $B_1 \subseteq B_2 \Rightarrow B'_2 \subseteq B'_1$
  - $A \subseteq A''$
  - $B \subseteq B''$
  - $A' = A'''$
  - $B' = B'''$
    - $A \subseteq B' \Leftrightarrow B \subseteq A' \Leftrightarrow A \times B \subseteq I$

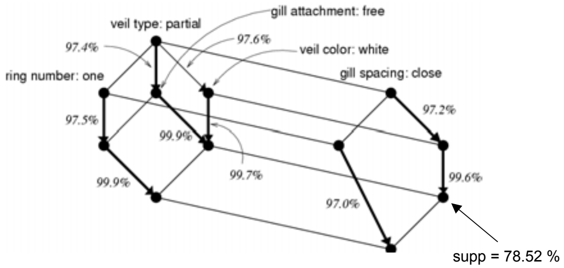
From $B' = B'''$ follows $\quad \text{supp}(B) = \frac{|B'|}{|G|} = \frac{|B'''|}{|G|} = \text{supp}(B'')$

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

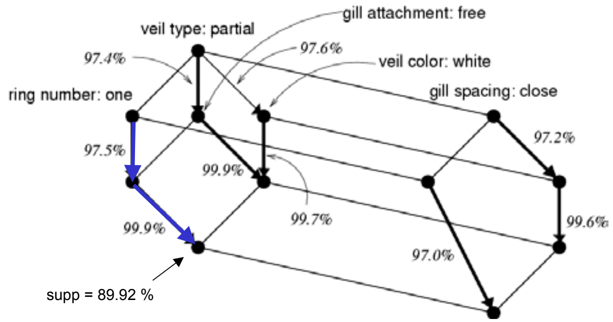


Association Rules can be visualized in the line diagram:
- exact rules as described before
- **approximate rules** as above ⑤ Luxenburger basis

Def.: The **Luxenburger basis** consists of all valid association rules $X \rightarrow Y$ such that there are concepts (A_1, B_1) and (A_2, B_2) where (A_1, B_1) is a direct upper neighbor of (A_2, B_2), $X = B_1$, and $X \cup Y = B_2$.



Each arrow indicates a rule of the basis, e.g. the rightmost arrow stands for {veil type: partial, gill spacing: close, veil color: white} → {gill attachment: free} (conf = 99.6 %, supp = 78.52 %)



All other rules can be derived:
For instance {ring number: one} → {veil color: white} has support 89.92 % (the support of the largest concept having both attributes in its intent) and confidence 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