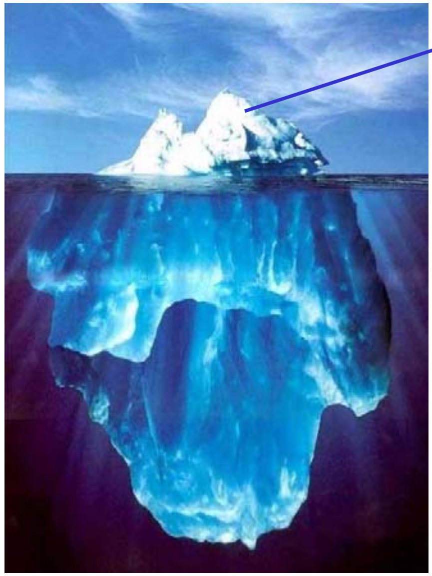


Efficient Data Mining Based on **Formal Concept Analysis**

Ergänzung zu Kap. 4 der KDD-Vorlesung SS 2005

Gerd Stumme

Institute for Applied Informatics (AIFB)

University of Karlsruhe, Germany

AIFB

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. Other Application(s) of FCA

AIFB

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} → {Hiking}

conf = 100 %, supp = 10/19

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

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

AIFB

Observation:

The rules

{ Boating } → { Hiking, NPS Guided Tours, Fishing }

{ Boating, Swimming } → { 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.

Conclusion:
It is sufficient to look at one of those sets!

→ **faster computation**

→ **no redundant rules**

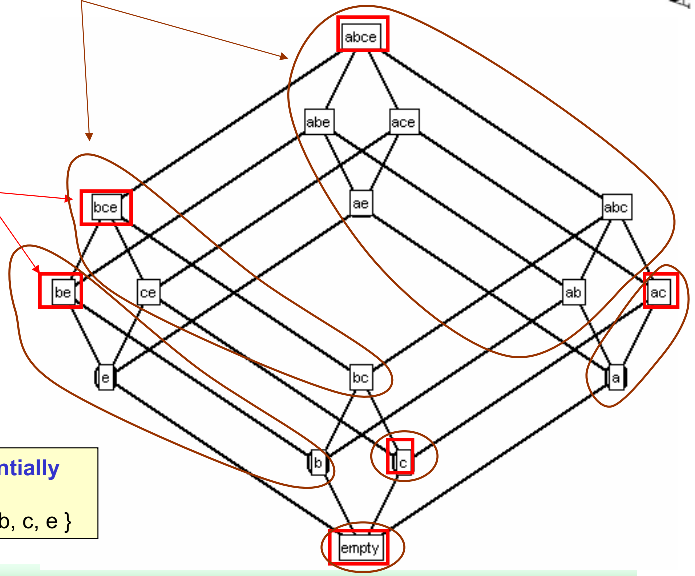
	NPS Guided Tours	Hiking	Horseback Riding	Swimming	Boating	Fishing	Bicycle Trail	Cross Country Trail
Cabrillo Natl. Mon.						×	×	
Channel Islands Natl. Park		×		×		×		
Death Valley Natl. Mon.	×	×	×	×			×	
Devils Postpile Natl. Mon.	×	×	×	×		×		
Fort Point Natl. Historic Site	×					×		
Golden Gate Natl. 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	×	×	×	×	×	×	×	×

AIFB

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 ∈ [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 $X \rightarrow Y$.

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 $X \rightarrow Y$.

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 $X \rightarrow Y$.

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 $X \rightarrow Y$.

Structure of the Talk:

- Introduction to FCA
- Conceptual Clustering with FCA
- Mining Association Rules with FCA
- Other Applications of FCA

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.

Structure of the Talk:

- Introduction to FCA
- Conceptual Clustering with FCA
- Mining Association Rules with FCA
- Other Applications of FCA

Two-Step Approach:

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

This is joint work with
L. Lakhal, Y. Bastide,
N. Pasquier, R. Taouil.

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. Other Application(s) of FCA

Formal Concept Analysis

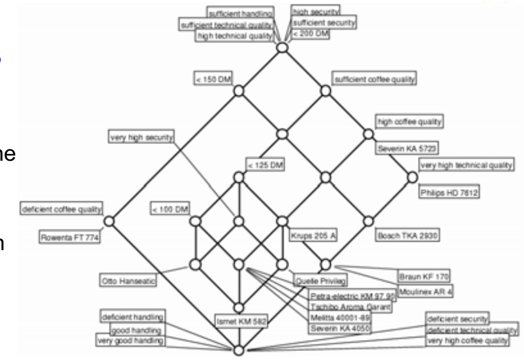
arose around 1980 in Darmstadt as a mathematical theory, which formalizes the concept of ‚concept'.

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.	Kaffee-qualität	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*.

- $(g,m) \in I$ 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	×	×	×	×	×	×	×	×

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	×	×	×	×	×	×	×	×

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´

A

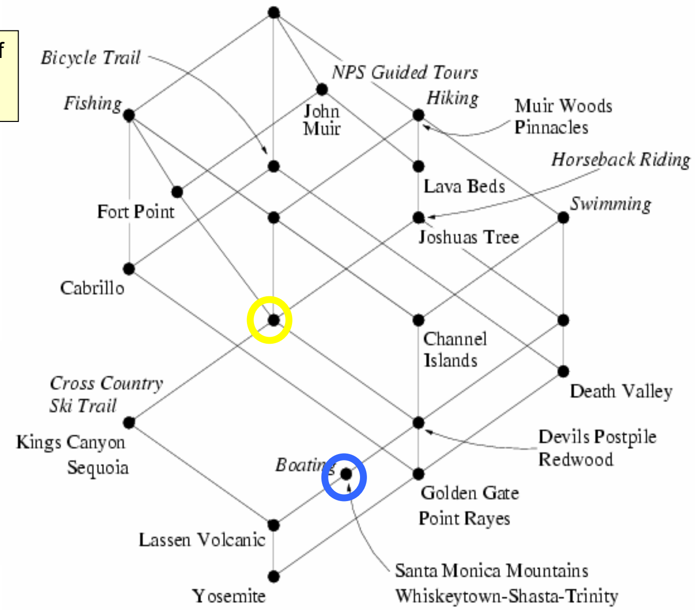
National Parks in California	NPS Guided Tours	Hiking	Horseback Riding	Swimming	Boating	Fishing	Bicycle Trail	Cross Country Trail
Cabrillo Natl. Mon.						×	×	
Channel Islands Natl. Park		×		×		×		
Death Valley Natl. Mon.	×	×	×	×			×	
Devils Postpile Natl. Mon.	×	×	×	×		×		
Fort Point Natl. Historic Site	×					×		
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	×	×	×	×	×	×	×	×

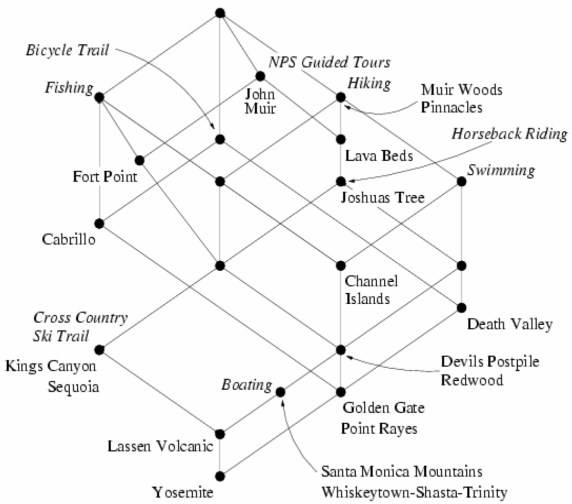
The **concept lattice** of the National Parks in California



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 }

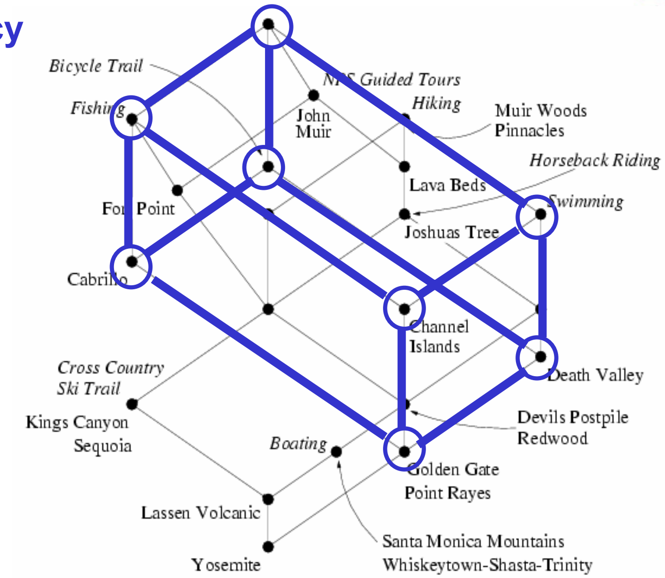
Independency

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

Example:

- Fishing
- Bicycle Trail
- Swimming

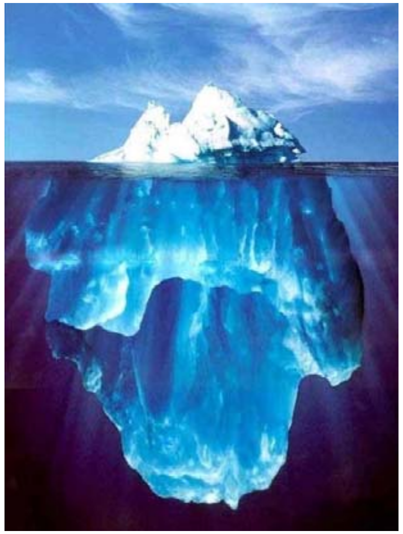
are independent attributes.

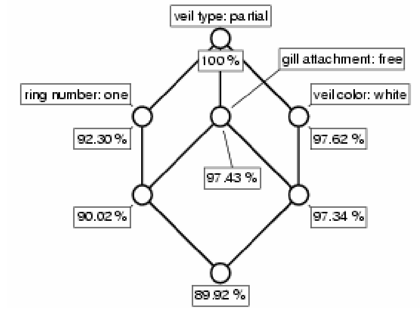


AIFB

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. Other Application(s) of FCA

Iceberg Concept Lattices

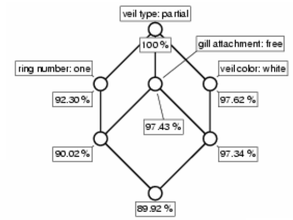




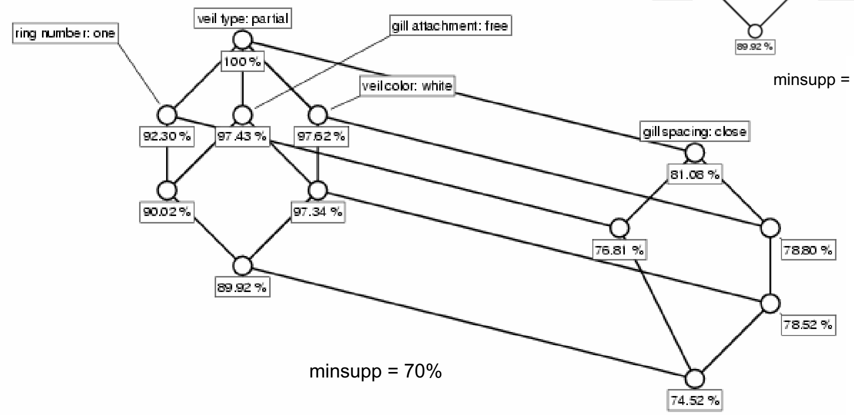
minsupp = 85%

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

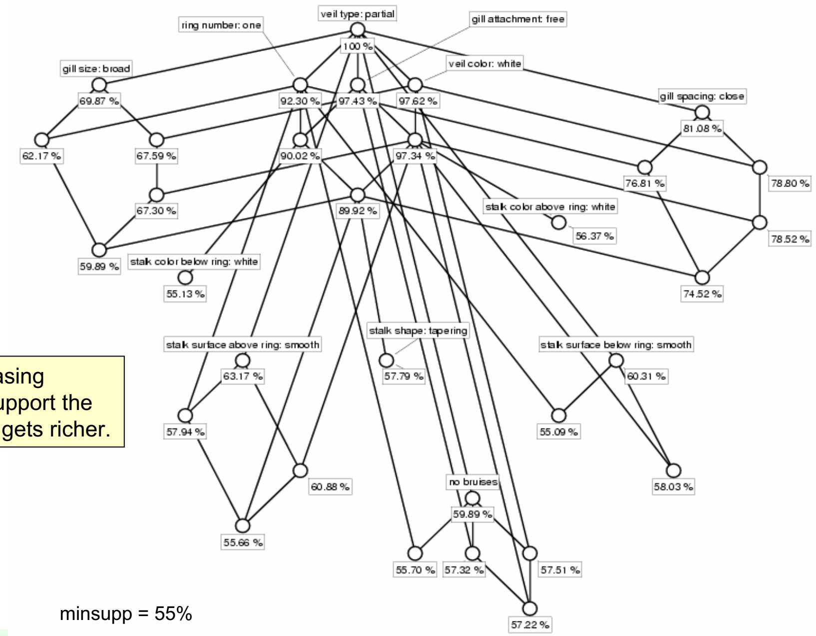
Iceberg Concept Lattices



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.

TITANIC

computes the iceberg concept lattice using the support:

Lemma 4. *Let* $X, Y \subseteq M$.

1. $X \subseteq Y \Longrightarrow \operatorname{supp}(X) \geq \operatorname{supp}(Y)$
2. $X'' = Y'' \Longrightarrow \operatorname{supp}(X) = \operatorname{supp}(Y)$
3. $X \subseteq Y \wedge \operatorname{supp}(X) = \operatorname{supp}(Y) \Longrightarrow X'' = Y''$

TITANIC

tries to optimize the following three questions:

1. How can the closure of an itemset be determined based on supports only?

2. How can the closure system be computed with determining as few closures as possible?

3. How can as many supports as possible be derived from already known supports?

TITANIC

1. How can the closure of an itemset be determined based on supports only?

X" = X ∪ { x∈ M \ X | supp(X) = supp(X ∪ x) }

Example: { b,c }" = { b, c, e }, since

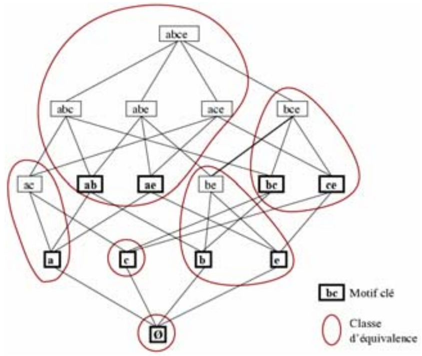
supp({ b, c }) = 1/3

and

supp({ a, b, c }) = 0/3

supp({ b, c, e }) = 1/3,

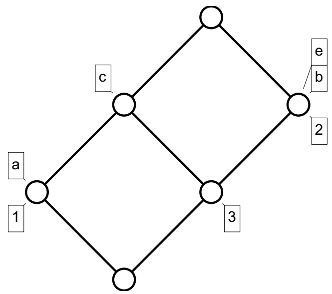
	a	b	c	e
1	×		×	
2		×		×
3		×	×	×

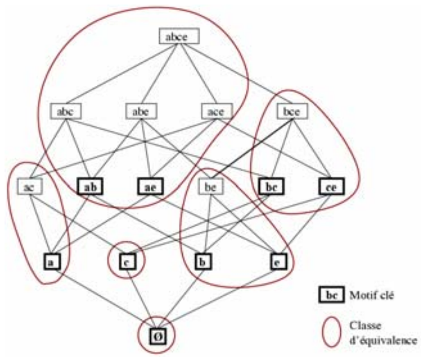


TITANIC

1. How can the closure of an itemset be determined based on supports only?

X" = X ∪ { x∈ M \ X | supp(X) = supp(X ∪ x) }





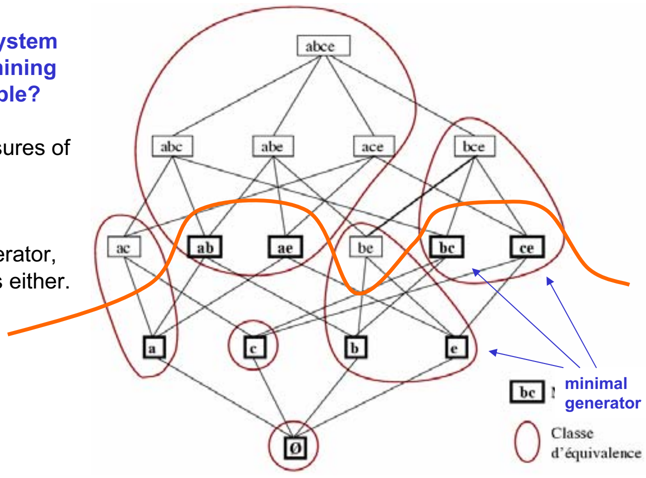
TITANIC

2. How can the closure system be computed with determining as few closures as possible?

We determine only the closures of the **minimal generators**.

• If a set is not minimal generator, then none of its supersets is either.

→ Apriori like approach



In the example, TITANIC needs two runs (and Apriori four).

TITANIC

1. How can the closure of an itemset be determined based on supports only?

$$X'' = X \cup \; x \in M \setminus X \mid supp(X) = supp(X \cup \; x)$$

2. How can the closure system be computed with determining as few closures as possible?

Approach à la Apriori

3. How can as many supports as possible be derived from already known supports?

3. How can as many supports as possible be derived from already known supports?

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

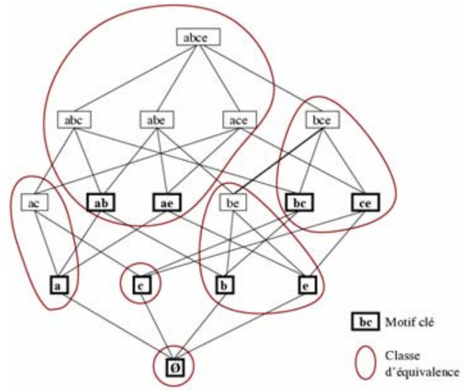
Theorem: If X is no minimal generator, then

supp(X) = min { supp(K) | K is minimal generator, $K \subseteq X$ } .

Example: supp({ a, b, c })

= min { supp({a, b }), supp({ b, c }), supp(a), supp(b), supp(c) }

= min { 0/3, 1/3, 1/3, 2/3, 2/3 } = 0,



TITANIC

1. How can the closure of an itemset be determined based on supports only?

$$X'' = X \cup \{ x \in M \setminus X \mid supp(X) = supp(X \cup x) \}$$

2. **How can the closure system be computed with determining as few closures as possible?**

Approach à la Apriori

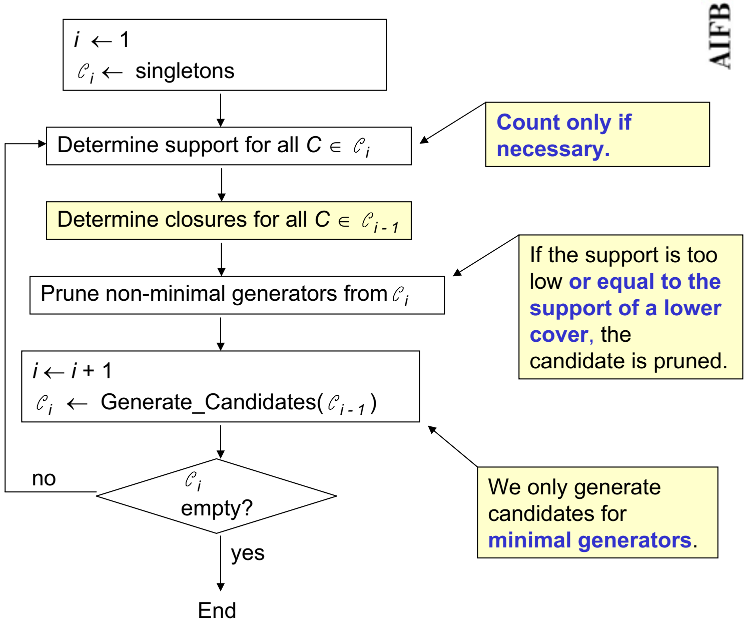
3. **How can as many supports as possible be derived from already known supports?**

If X is no minimal generator, then

supp(X) = min { supp(K) | K is minimal generator, $K \subseteq X$ } .

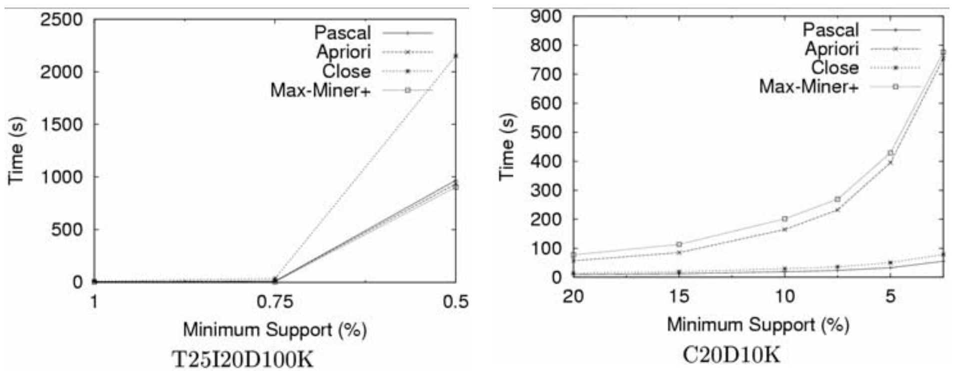
TITANIC

compared with Apriori



Pascal/Titanic

compared with Apriori

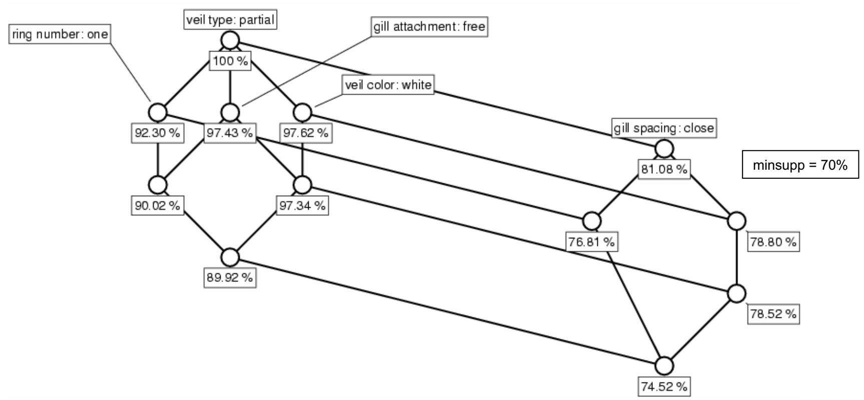


Weakly correlated data: similar performance of Pascal, Apriori and Max-Miner

Strongly correlated data: Pascal (and Close) are very efficient

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. Other Application(s) of FCA

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 $\text{supp}(B) = \text{supp}\,(B'')$ follows:

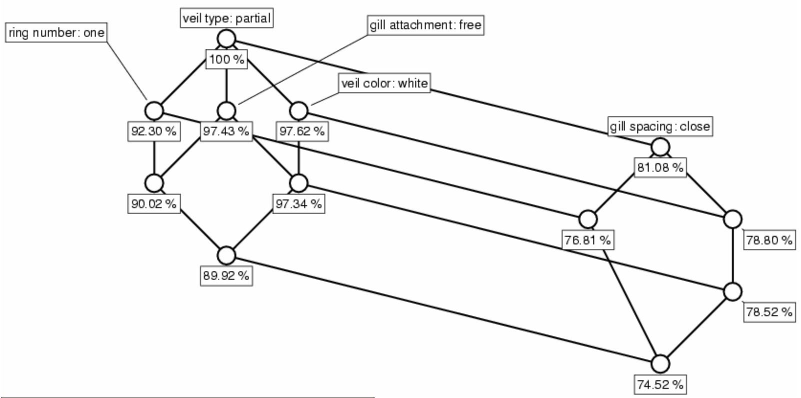
Theorem: $X \rightarrow Y$ and $X'' \rightarrow Y''$ have the same support and the same confidence.

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** ← conf = 100 %
- **approximate rules** ← conf < 100 %

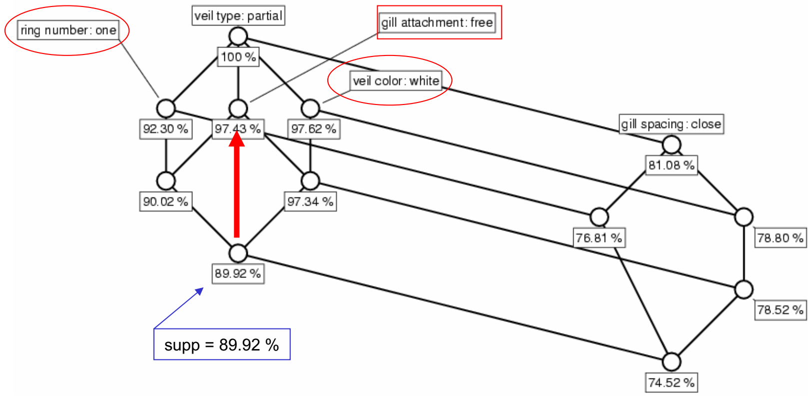
Exact association rules



Association rules can be visualized in the iceberg concept lattice:

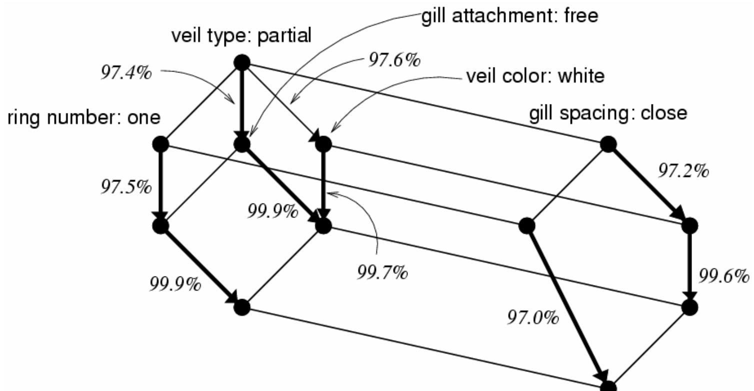
- **exact rules** ← conf = 100 %
- **approximate rules** ← conf < 100 %

Exact association rules



{ring number: one, veil color: white} → {gill attachment: free}
supp = 89.92 % conf = 100 %.

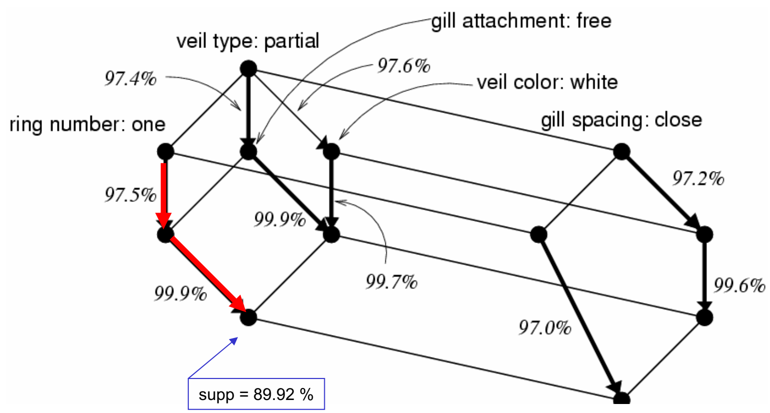
Luxenburger Basis for approximate association rules



Association rules can be visualized in the iceberg concept lattice:

- **exact rules** ← conf = 100 %
- **approximate rules** ← conf < 100 %

Luxenburger Basis for approximate association rules



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

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

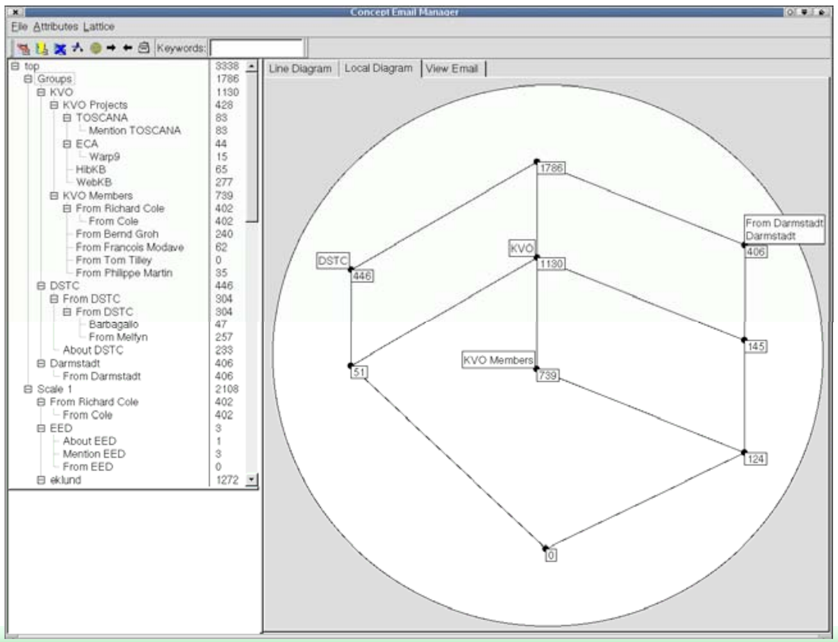
Name	Number of objects	Average size of objects	Number of items
T10I4D100K	100,000	10	1,000
MUSHROOMS	8,416	23	127
C20D10K	10,000	20	386
C73D10K	10,000	73	2,177

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. **Other Application(s) of FCA**

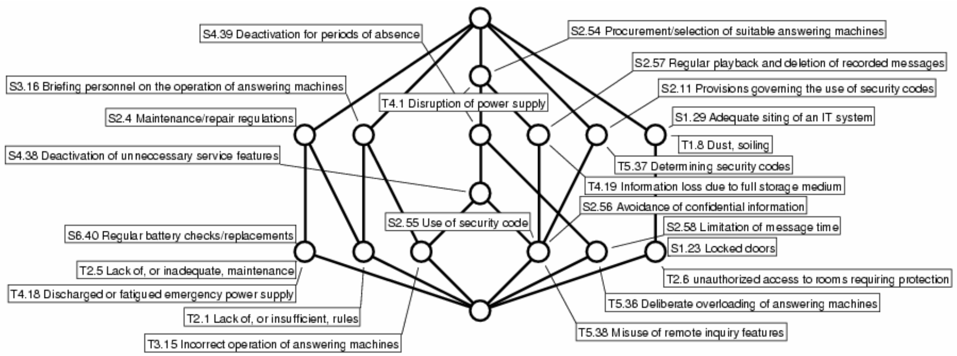
Conceptual Email Manager



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. **Other Application(s) of FCA**

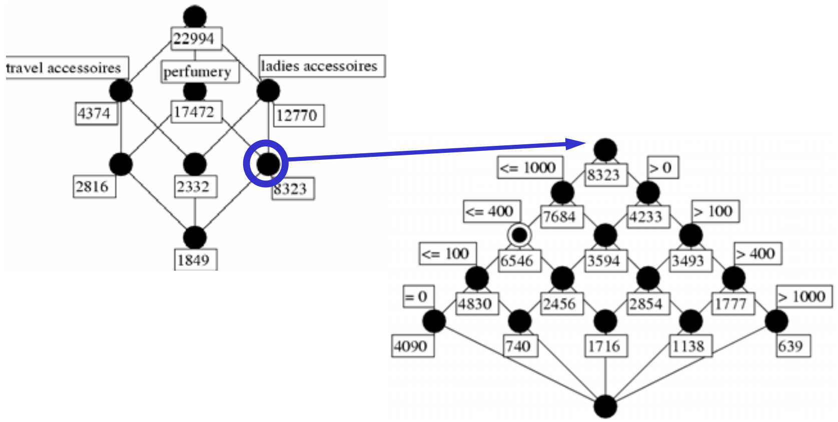
IT-Security Management

- Supports the analysis of security risks in IT units
- status quo test for establishing guidelines and checklists



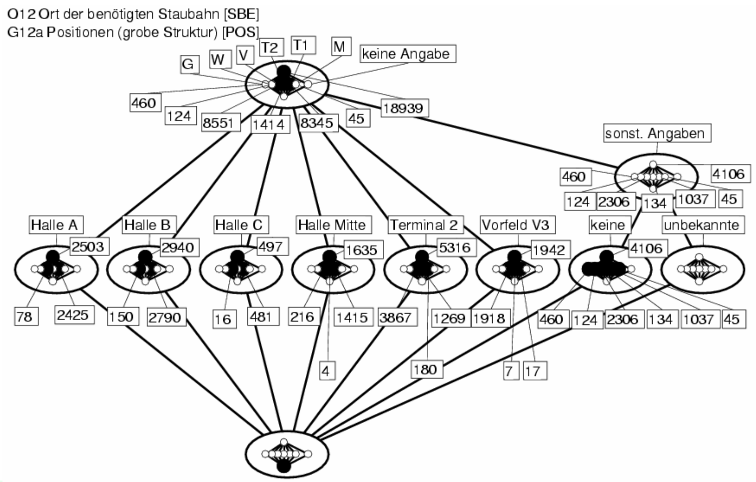
Database Marketing at Jelmoli AG, Zürich

- Analysis of the user behavior of customers using the Shopping Bonus Card
- Supporting of Cross-Selling via Direct Mailing

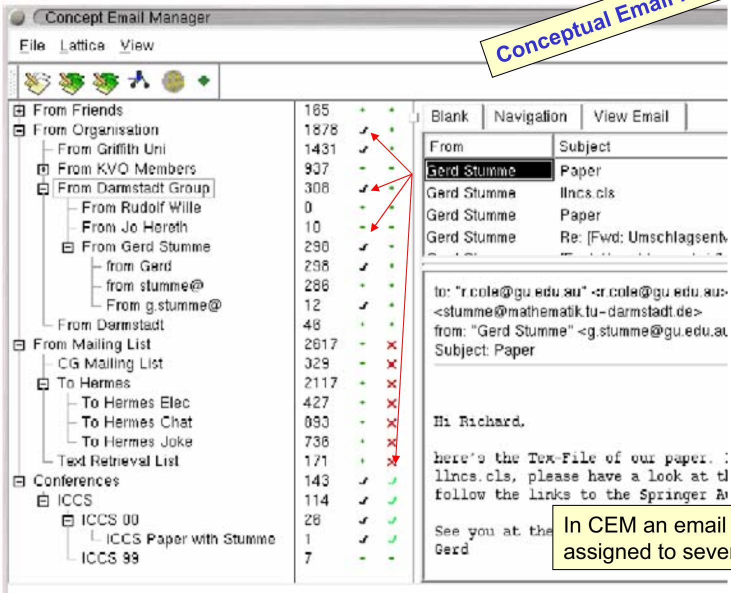


Analysis of flight movements at Frankfurt Airport

- Allowing for ad-hoc queries in the database
- Visualization of dependencies

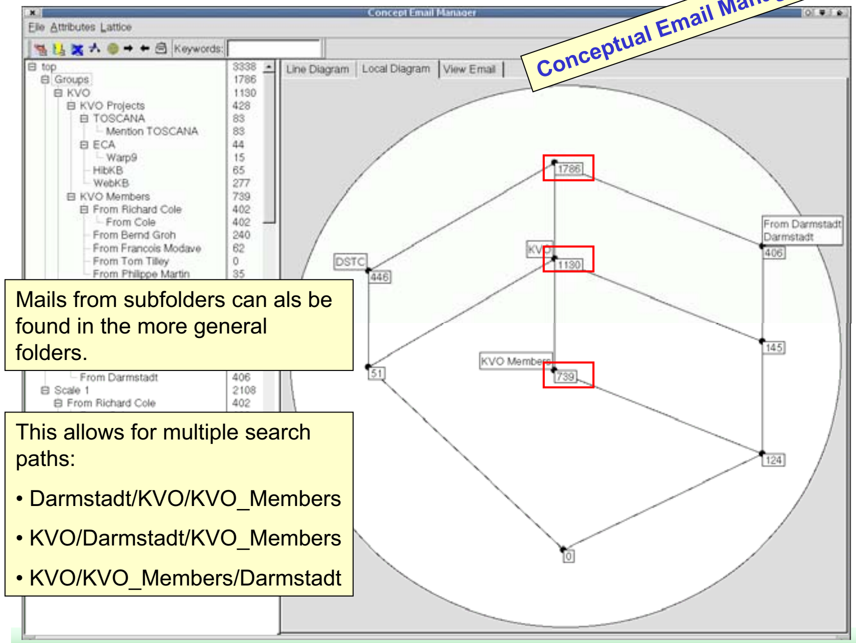


Conceptual Email Manager



In CEM an email can be assigned to several „folders“.

Conceptual Email Manager

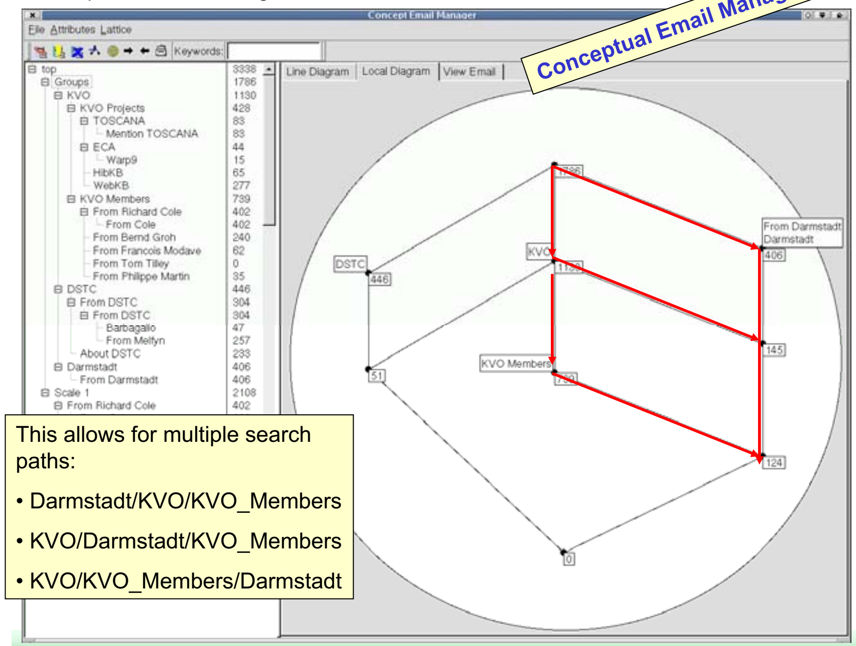


Mails from subfolders can als be found in the more general folders.

This allows for multiple search paths:

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

Conceptual Email Manager



This allows for multiple search paths:

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

Conceptual Email Manager

Nested line diagrams allow the combination of views.