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
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**
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?

\{Swimming\} → \{Hiking\}

conf = 100 %, supp = 10/19

#(swimming+hiking parks) / #(swimming parks)
Observation:
The rules

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

have the same support and the same confidence, because the two sets
\{ \text{Boating} \} and \{ \text{Boating, Swimming} \}
describe exactly the same set of parks.

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

$\rightarrow$ faster computation
$\rightarrow$ no redundant rules
Another Toy Example:

Unique representatives 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.

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

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

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).
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
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 $X \rightarrow Y$.

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).

Structure of the Talk:
- Introduction to FCA
- Conceptual Clustering with FCA
- Mining Association Rules with FCA
- Other Applications of FCA

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.
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 esthetics
- 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 \(m\)“.

<table>
<thead>
<tr>
<th>National Parks in California</th>
<th>NPS Guided Tours</th>
<th>Hiking</th>
<th>Horseback Riding</th>
<th>Swimming</th>
<th>Boating</th>
<th>Fishing</th>
<th>Bicycle Trail</th>
<th>Cross Country Trail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cabrillo Natl. Mon.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Channel Islands Natl. Park</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Death Valley Natl. Mon.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Devils Postpile Natl. Mon.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fort Point Natl. Historic Site</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Golden Gate Natl. Recreation Area</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>John Muir Natl. Historic Site</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Joshua Tree Natl. Mon.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kings Canyon Natl. Park</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lassen Volcanic Natl. Park</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lava Beds Natl. Mon.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Muir Woods Natl. Mon.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pinnacles Natl. Mon.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point Reyes Natl. Seashore</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Redwood Natl. Park</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Santa Monica Mts. Natl. Recr. Area</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequoia Natl. Park</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whiskeytown-Shasta-Trinity Natl. Recr. Area</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yosemite Natl. Park</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
For $A \subseteq G$, we define

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

For $B \subseteq M$, we define dually

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

### National Parks in California

<table>
<thead>
<tr>
<th>National Park</th>
<th>NPS Guided Tours</th>
<th>Hiking</th>
<th>Horseback Riding</th>
<th>Swimming</th>
<th>Boating</th>
<th>Fishing</th>
<th>Bicycle Trail</th>
<th>Cross Country Trail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cabrillo Natl. Mon.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Channel Islands Natl. Park</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Death Valley Natl. Mon.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Devils Postpile Natl. Mon.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fort Point Natl. Historic Site</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Golden Gate Natl. Recreation Area</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>John Muir Natl. Historic Site</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Joshua Tree Natl. Mon.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kings Canyon Natl. Park</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lassen Volcanic Natl. Park</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lava Beds Natl. Mon.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Muir Woods Natl. Mon.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pinnacles Natl. Mon.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point Reyes Natl. Seashore</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Redwood Natl. Park</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Santa Monica Mts. Natl. Recr. Area</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequoia Natl. Park</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whiskeytown-Shasta-Trinity Natl. Recr. Area</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yosemite Natl. Park</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**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 \(B' = A\).

\[
\begin{align*}
\text{Cabrillo Natl. Mon.} & & & & \times & \times \\
\text{Channel Islands Natl. Park} & & \times & \times & \times \\
\text{Death Valley Natl. Mon.} & & \times & \times & \times & \times \\
\text{Devils Postpile Natl. Mon.} & & \times & \times & \times & \times \\
\text{Fort Point Natl. Historic Site} & & \times & \times & \times \\
\text{Golden Gate Natl. Recreation Area} & & \times & \times & \times & \times & \times \\
\text{John Muir Natl. Historic Site} & & \times \\
\text{Joshua Tree Natl. Mon.} & & \times & \times & \times \\
\text{Kings Canyon Natl. Park} & & \times & \times & \times & \times & \times \\
\text{Lassen Volcanic Natl. Park} & & \times & \times & \times & \times & \times & \times \\
\text{Lava Beds Natl. Mon.} & & \times & \times & \times \\
\text{Muir Woods Natl. Mon.} & & \times \\
\text{Pinnacles Natl. Mon.} & & \times \\
\text{Point Reyes Natl. Seashore} & & \times & \times & \times & \times & \times \\
\text{Redwood Natl. Park} & & \times & \times & \times & \times \\
\text{Santa Monica Mts. Natl. Recr. Area} & & \times & \times & \times & \times & \times \\
\text{Sequoia Natl. Park} & & \times & \times & \times & \times & \times \\
\text{Whiskeytown-Shasta-Trinity Natl. Recr. Area} & & \times & \times & \times & \times & \times & \times \\
\text{Yosemite Natl. Park} & & \times & \times & \times & \times & \times & \times & \times
\end{align*}
\]
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.)

<table>
<thead>
<tr>
<th>National Parks in California</th>
<th>NPS Guided Tours</th>
<th>Horseback Riding</th>
<th>Swimming</th>
<th>Boating</th>
<th>Fishing</th>
<th>Bicycle Trail</th>
<th>Cross Country Trail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cabrillo Natl. Mon.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Channel Islands Natl. Park</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Death Valley Natl. Mon.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Devils Postpile Natl. Mon.</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fort Point Natl. Historic Site</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Golden Gate Natl. Recreation Area</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>John Muir Natl. Historic Site</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Joshua Tree Natl. Mon.</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kings Canyon Natl. Park</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lassen Volcanic Natl. Park</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lava Beds Natl. Mon.</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Muir Woods Natl. Mon.</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pinnacles Natl. Mon.</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point Reyes Natl. Seashore</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Redwood Natl. Park</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Santa Monica Mts. Natl. Recr. Area</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequoia Natl. Park</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whiskeytown-Shasta-Trinity Natl. Recr. Area</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yosemite Natl. Park</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The **concept lattice** of the National Parks in California

![Diagram of 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 }
Attributes are independent if they span a hyper-cube (i.e., if all $2^n$ combinations 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. Other Application(s) of FCA
For \( \text{minsupp} = 85\% \) 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.

\[ \text{minsupp} = 55\% \]
Iceberg Concept Lattices and Frequent Itemsets

Iceberg concept lattices are a condensed representation of frequent itemsets:

\[ \text{supp}(X) = \text{supp}(X'') \]

<table>
<thead>
<tr>
<th>minsupp</th>
<th># frequent closed itemsets</th>
<th># frequent itemsets</th>
</tr>
</thead>
<tbody>
<tr>
<td>85 %</td>
<td>7</td>
<td>16</td>
</tr>
<tr>
<td>70 %</td>
<td>12</td>
<td>32</td>
</tr>
<tr>
<td>55 %</td>
<td>32</td>
<td>116</td>
</tr>
<tr>
<td>0 %</td>
<td>32.086</td>
<td>(2^{80})</td>
</tr>
</tbody>
</table>

Difference between frequent concepts and frequent itemsets in the mushrooms database.
computes the iceberg concept lattice using the support:

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

1. $X \subseteq Y \implies \text{supp}(X) \geq \text{supp}(Y)$
2. $X'' = Y'' \implies \text{supp}(X) = \text{supp}(Y)$
3. $X \subseteq Y \land \text{supp}(X) = \text{supp}(Y) \implies X'' = Y''$
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?
1. How can the closure of an itemset be determined based on supports only?

\[ X'' = X \cup \{ x \in M \setminus X | \text{supp}(X) = \text{supp}(X \cup x) \} \]

Example: \( \{ b, c \}'' = \{ b, c, e \} \), since

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

and

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

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

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

\[ X'' = X \cup \{ x \in M \setminus X | \text{supp}(X) = \text{supp}(X \cup x) \} \]

Example: \( \{ b, c \}'' = \{ b, c, e \} \), since

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

and

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

\( \text{supp}( \{ b, c, e \} ) = 1/3, \)
1. How can the closure of an itemset be determined based on supports only?

\[ X^\prime = X \cup \{ x \in M \setminus X \mid \text{supp}(X) = \text{supp}(X \cup x) \} \]
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).
1. How can the closure of an itemset be determined based on supports only?

\[
X^\prime\prime = X \cup \{ x \in M \setminus X \mid \text{supp}(X) = \text{supp}(X \cup x) \}
\]

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?

**Theorem:** If $X$ is no minimal generator, then

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

**Example:** $\text{supp}(\{a, b, c\})$

$$= \min \{ \text{supp}(\{a, b\}), \text{supp}(\{b, c\}), \text{supp}(a), \text{supp}(b), \text{supp}(c) \}$$

$$= \min \{ 0/3, 1/3, 1/3, 2/3, 2/3 \} = 0,$$
1. How can the closure of an itemset be determined based on supports only?

\[ X^\prime = X \cup \{ x \in M \setminus X \mid \text{supp}(X) = \text{supp}(X \cup x) \} \]

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

\[ \text{supp}(X) = \min \{ \text{supp}(K) \mid K \text{ is minimal generator, } K \subseteq X \} . \]
We only generate candidates for minimal generators.

\[ \begin{align*}
    i & \leftarrow 1 \\
    \mathcal{C}_i & \leftarrow \text{singletons} \\
    \text{Determine support for all } C \in \mathcal{C}_i \\
    \text{Determine closures for all } C \in \mathcal{C}_{i-1} \\
    \text{Prune non-minimal generators from } \mathcal{C}_i \\
    i & \leftarrow i + 1 \\
    \mathcal{C}_i & \leftarrow \text{Generate}_\text{Candidates}(\mathcal{C}_{i-1}) \\
    \mathcal{C}_i & \text{ empty?} \\
    \text{no} \\
    \text{yes} \\
    \text{End}
\end{align*} \]
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!)

minsupp = 70%
• 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**
- **approximate rules**

$\text{conf} = 100\%$

$\text{conf} < 100\%$
Association rules can be visualized in the iceberg concept lattice:

- **exact rules**
  - conf = 100 %

- **approximate rules**
  - conf < 100 %
Exact association rules

\{\text{ring number: one, veil color: white}\} \rightarrow \{\text{gill attachment: free}\}

supp = 89.92 % \quad \text{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%.
Some experimental results

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

© Gerd Stumme 2002     Invited Talk at DEXA '2
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

The End
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
In CEM an email can be assigned to several "folders".
Conceptual Email Manager

This allows for multiple search paths:

- Darmstadt/KVO/KVO_Members
- KVO/Darmstadt/KVO_Members
- KVO/KVO_Members/Darmstadt
Mails from subfolders can also 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
Nested line diagrams allow the combination of views.