Effizientes Data Mining mit Formaler Begriffsanalyse

Vorlesung Knowledge Discovery
Kap. 9
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?

\{Swimming\} → \{Hiking\}

conf = 100 %, supp = 10/19

\frac{\#(swimming+hiking parks)}{\#(swimming parks)} / \frac{\#(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.

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

→ faster computation
→ 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 \( \in [0,1] \), 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 $X \rightarrow Y$.
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
- Frequent (Closed) Datalog Queries

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

For $B \subseteq M$, we define dually
$$B^\prime := \{ g \in G \mid \forall m \in B: (g,m) \in I \}.$$ 

<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>X</td>
<td></td>
</tr>
<tr>
<td>Channel Islands Natl. Park</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</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 \(B' = A\).

---

**National Parks in California**

<table>
<thead>
<tr>
<th>National Park</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>
</tr>
<tr>
<td>Channel Islands Natl. Park</td>
<td>(\times)</td>
<td>(\times)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Death Valley Natl. Mon.</td>
<td>(\times)</td>
<td>(\times)</td>
<td>(\times)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Devils Postpile Natl. Mon.</td>
<td>(\times)</td>
<td>(\times)</td>
<td>(\times)</td>
<td>(\times)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fort Point Natl. Historic Site</td>
<td>(\times)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Golden Gate Natl. Recreation Area</td>
<td>(\times)</td>
<td>(\times)</td>
<td>(\times)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>John Muir Natl. Historic Site</td>
<td>(\times)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Joshua Tree Natl. Mon.</td>
<td>(\times)</td>
<td>(\times)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kings Canyon Natl. Park</td>
<td>(\times)</td>
<td>(\times)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lassen Volcanic Natl. Park</td>
<td>(\times)</td>
<td>(\times)</td>
<td>(\times)</td>
<td>(\times)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lava Beds Natl. Mon.</td>
<td>(\times)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Muir Woods Natl. Mon.</td>
<td>(\times)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pinnacles Natl. Mon.</td>
<td>(\times)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point Reyes Natl. Seashore</td>
<td>(\times)</td>
<td>(\times)</td>
<td>(\times)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Redwood Natl. Park</td>
<td>(\times)</td>
<td>(\times)</td>
<td>(\times)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Santa Monica Mts. Natl. Recr. Area</td>
<td>(\times)</td>
<td>(\times)</td>
<td>(\times)</td>
<td>(\times)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequoia Natl. Park</td>
<td>(\times)</td>
<td>(\times)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whiskeytown-Shasta-Trinity Natl. Recr. Area</td>
<td>(\times)</td>
<td>(\times)</td>
<td>(\times)</td>
<td>(\times)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yosemite Natl. Park</td>
<td>(\times)</td>
<td>(\times)</td>
<td>(\times)</td>
<td>(\times)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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>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>
The **concept lattice** of the National Parks in California

<table>
<thead>
<tr>
<th>National Parks in California</th>
<th>Bike</th>
<th>Hike</th>
<th>Cross Country Ski Trail</th>
<th>Fishing</th>
<th>Fort Point</th>
<th>Boat</th>
<th>NPS Guided Tours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kings Canyon Natl. Park</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lassen Volcanic Natl. Park</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>
<tr>
<td>Point Reyes Natl. Seashore</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Additional outdoor activities include:
- **Horseback Riding**
- **Swimming**
- **Devils Postpile**
- **Point Reyes**
- **Santa Monica Mountains**
- **Whiskeytown-Shasta-Trinity National Recreation Area**
**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. Text Clustering with Background Knowledge
Iceberg Concept Lattices

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

\[ \text{minsupp} = 85\% \]

\[ \text{minsupp} = 70\% \]
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:

\[ \text{supp}(X) = \text{supp}(X^{\prime}) \]

<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.
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 \( \text{supp}(B) = \text{supp}(B^{\prime \prime}) \) follows:

**Theorem:** \( X \rightarrow Y \) and \( X^{\prime \prime} \rightarrow Y^{\prime \prime} \) 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 \( B = B^{\prime \prime} \) (i.e., the intents of the iceberg concept lattice).

Association rules can be visualized in the iceberg concept lattice:

- **exact rules**
  - \( \text{conf} = 100 \% \)
- **approximate rules**
  - \( \text{conf} < 100 \% \)
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
- approximate rules

conf = 100 %
conf < 100 %
Luxenburger Basis for approximate association rules

\{\text{ring number: one}\} \rightarrow \{\text{veil color: white}\}

\text{supp} = 89.92\% \quad \text{conf} = 97.5\% \times 99.9\% \approx 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>70%</td>
<td></td>
<td>20,419</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td></td>
<td>21,686</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>30%</td>
<td></td>
<td>22,952</td>
<td></td>
<td></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>70%</td>
<td></td>
<td>37,671</td>
<td></td>
<td>968</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td></td>
<td>56,703</td>
<td></td>
<td>1,169</td>
</tr>
<tr>
<td></td>
<td>30%</td>
<td></td>
<td>71,412</td>
<td></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>70%</td>
<td></td>
<td>89,601</td>
<td></td>
<td>1,948</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td></td>
<td>116,791</td>
<td></td>
<td>1,948</td>
</tr>
<tr>
<td></td>
<td>30%</td>
<td></td>
<td>116,791</td>
<td></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>90%</td>
<td></td>
<td>2,053,896</td>
<td></td>
<td>4,089</td>
</tr>
<tr>
<td></td>
<td>85%</td>
<td></td>
<td>2,053,936</td>
<td></td>
<td>4,089</td>
</tr>
<tr>
<td></td>
<td>80%</td>
<td></td>
<td>2,053,936</td>
<td></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. 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

Clusterberechnung

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?
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

- choose a representation
- similarity measure
- clustering algorithm

- Reuters data set for our studies (min15 max100)
- bag of terms (details on the next slide)
- cosine as similarity measure
- Bi-Section is a version of KMeans
- Conceptual clustering with FCA
Preprocessing steps

- build a bag of words model

<table>
<thead>
<tr>
<th>docid</th>
<th>term1</th>
<th>term2</th>
<th>term3</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>doc1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>doc2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>doc3</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>doc4</td>
<td>2</td>
<td>23</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

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

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

| \( |D| \) | 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 Terms

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

4. Generalizations:

- 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

- consists of ‘Synsets’, which group synonyms together.
- the synsets are hierarchically organized.
- is online under http://wordnet.princeton.edu
Hinzufügen von Oberbegriffen aus WordNet

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

Strategies: all, first, context

109377 Concepts (synsets)

144684 lexical entries

EN: oil
EN: anoint
EN: inunct
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.
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 dlr per barrel adopted by non-OPEC Oman since February. REUTERS
• 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 $E_1, E_2$
    - 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 $\theta$ liegen

**Begriffliches Clustern mit Begriffsanalyse**

- Berechnung des Begriffsverbandes erzeugt intensionale Beschreibungen der Cluster
- Visualisierung
### Extracted Word Description

<table>
<thead>
<tr>
<th>Cluster 0</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Cluster 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>amount</td>
<td>0.12</td>
<td>depository financial institution</td>
<td>0.09</td>
<td>loss</td>
</tr>
<tr>
<td>billion, one million million</td>
<td>0.11</td>
<td>financial institution, finance</td>
<td>0.09</td>
<td>failure</td>
</tr>
<tr>
<td>integer, whole number'</td>
<td>0.11</td>
<td>rate, charge per unit'</td>
<td>0.09</td>
<td>nonaccomplishment, nono</td>
</tr>
<tr>
<td>insufficiency, inadequacy</td>
<td>0.11</td>
<td>institution, establishment</td>
<td>0.09</td>
<td>ten, 10, X, ten, decade</td>
</tr>
<tr>
<td>deficit, shortage, shortfall</td>
<td>0.11</td>
<td>loss</td>
<td>0.08</td>
<td>American state'</td>
</tr>
<tr>
<td>number</td>
<td>0.09</td>
<td>monetary unit'</td>
<td>0.07</td>
<td>state, province'</td>
</tr>
<tr>
<td>excess, surplus, surplus</td>
<td>0.09</td>
<td>central, telephone exchange</td>
<td>0.07</td>
<td>system, unit'</td>
</tr>
<tr>
<td>overabundance, overmuch</td>
<td>0.09</td>
<td>financial loss'</td>
<td>0.06</td>
<td>network, net, mesh, mesh</td>
</tr>
<tr>
<td>abundance, copiousness</td>
<td>0.09</td>
<td>outgo, expenditure, outlay</td>
<td>0.06</td>
<td>September, Sep, Sept'</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster 5</th>
<th>Cluster 6</th>
<th>Cluster 7</th>
<th>Cluster 8</th>
<th>Cluster 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>text, textual matter'</td>
<td>0.15</td>
<td>loss</td>
<td>0.34</td>
<td>gross sales, gross revenue</td>
</tr>
<tr>
<td>matter</td>
<td>0.15</td>
<td>failure</td>
<td>0.33</td>
<td>sum, sum of money, amd</td>
</tr>
<tr>
<td>letter, missive'</td>
<td>0.15</td>
<td>nonaccomplishment, nono</td>
<td>0.32</td>
<td>income</td>
</tr>
<tr>
<td>sign, mark'</td>
<td>0.13</td>
<td>common fraction, simple</td>
<td>0.02</td>
<td>financial gain'</td>
</tr>
<tr>
<td>clue, clew, cue'</td>
<td>0.13</td>
<td>fraction</td>
<td>0.22</td>
<td>gain</td>
</tr>
<tr>
<td>purpose, intent, intention</td>
<td>0.11</td>
<td>rational number'</td>
<td>0.22</td>
<td>enterprise</td>
</tr>
<tr>
<td>evidence</td>
<td>0.11</td>
<td>real number, real'</td>
<td>0.22</td>
<td>business, concern, business</td>
</tr>
<tr>
<td>indication, indicant'</td>
<td>0.11</td>
<td>complex number, complex</td>
<td>0.22</td>
<td>assets</td>
</tr>
<tr>
<td>goal, end'</td>
<td>0.1</td>
<td>one-half, half'</td>
<td>0.22</td>
<td>division</td>
</tr>
<tr>
<td>writing, written material, n</td>
<td>0.07</td>
<td>revolutions per minute, rpm</td>
<td>0.22</td>
<td>army unit'</td>
</tr>
</tbody>
</table>

*Slide 52*
Ergebnisse

Begriffskette mit zunehmender Spezifität

refiner

oil

compound, chemical compound
Ergebnisse

Crude oil barrel
Ergebnisse

- Petroleums, crude oil, crude, coal oil, roc
- Tube, tubing
- Conduit, channel
- Barrel, gun barrel

- Resin
- Palm
Literatur


• Andreas Hotho, Steffen Staab, Gerd Stumme: *WordNet improves text document clustering*; Semantic Web Workshop @ SIGIR 2003.


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

The End