Machine Learning: Pattern Mining

Steffen Rendle

Information Systems and Machine Learning Lab (ISMLL)
University of Hildesheim

Wintersemester 2007 / 2008
Pattern Mining

Overview

Itemsets

Task
Naive Algorithm
Apriori Algorithm
Data Structure
Eclat Algorithm

Association Rules

Task
Algorithm

Summary
Overview

Pattern Mining discovers regularities in data.

- Example: a transaction database of a supermarket: someone who buys chips also buys beer.
- Frequent patterns are found by counting the occurrences in the data base.
- Types of patterns: itemsets, association rules, sequences, ...
## Example

<table>
<thead>
<tr>
<th>Shopping Carts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beer, Chips, Chocolate, Cookies</td>
</tr>
<tr>
<td>Coke, Beer, Pizza, Chips</td>
</tr>
<tr>
<td>Salad, Noodles, Tomatoes, Water</td>
</tr>
<tr>
<td>Lasagne, Coke, Beer, Chips</td>
</tr>
<tr>
<td>Oranges, Apple Juice, Rice, Cabbage, Sausage</td>
</tr>
<tr>
<td>Diapers, Beer, Charcoal, Sausage</td>
</tr>
<tr>
<td>Beer, Cabbage, Sausage, Chips</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>
### Example

**Shopping Carts**

<table>
<thead>
<tr>
<th>Beer, Chips, Chocolate, Cookies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coke, Beer, Pizza, Chips</td>
</tr>
<tr>
<td>Salad, Noodles, Tomatoes, Water</td>
</tr>
<tr>
<td>Lasagne, Coke, Beer, Chips</td>
</tr>
<tr>
<td>Oranges, Apple Juice, Rice, Cabbage, Sausage</td>
</tr>
<tr>
<td>Diapers, Beer, Charcoal, Sausage</td>
</tr>
<tr>
<td>Beer, Cabbage, Sausage, Chips</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

**Observations:**

- Many customers buy beer.
- Beer and chips are often bought together.
- Customers who buy cabbage also buy sausage.
- Customers who buy something to eat also buy something to drink.
Outline

- **Classification** predicts class labels based on training data
- **Clustering** groups data based on similarity
- **Pattern Mining** discovers regularities in data
Itemsets

- Which itemsets frequently occur in the same transaction?

- Example: chips and beer are frequently bought together

- given
  - Items \( I = \{i_1, \ldots, i_m\} \)
  - Data \( D \subseteq \mathcal{P}(I) \) multiset
  - Frequency threshold \( \theta_s \)

- to find
  - Frequent sets \( L = \{X \in \mathcal{P}(I)|support_D(X) \geq \theta_s\} \)
Definitions and Terms

- \( \text{support}_D(X) = \frac{|\{d \in D | X \subseteq d\}|}{|D|} \)

- \( X \) is frequent / large iff \( \text{support}_D(X) \geq \theta_s \)
Naive Algorithm

\begin{verbatim}
function \texttt{Naive}(D, \theta_s)
    \textbf{let} L \leftarrow \emptyset
    \textbf{for all} \ X \in \mathcal{P}(I) \ \textbf{do}
        \textbf{if} support_D(X) \geq \theta_s \ \textbf{then}
            \textbf{let} L \leftarrow L \cup \{X\}
        \textbf{end if}
    \textbf{end for}
    \textbf{return} L
end function
\end{verbatim}
## Example

<table>
<thead>
<tr>
<th>Data $D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>a,b,e</td>
</tr>
<tr>
<td>b,c</td>
</tr>
<tr>
<td>a,c,e</td>
</tr>
<tr>
<td>a,b,c,e</td>
</tr>
<tr>
<td>a,b,d,e</td>
</tr>
<tr>
<td>b,c,d</td>
</tr>
<tr>
<td>a,b,c</td>
</tr>
<tr>
<td>a,c</td>
</tr>
<tr>
<td>a,b,e</td>
</tr>
</tbody>
</table>

Find itemsets with $\theta_s \geq 0.3$

$X$ frequent $\iff \#_D(X) > 2$
## Example

### Data $D$

<table>
<thead>
<tr>
<th>Itemset</th>
</tr>
</thead>
<tbody>
<tr>
<td>a,b,e</td>
</tr>
<tr>
<td>b,c</td>
</tr>
<tr>
<td>a,c,e</td>
</tr>
<tr>
<td>a,b,c,e</td>
</tr>
<tr>
<td>a,b,d,e</td>
</tr>
<tr>
<td>b,c,d</td>
</tr>
<tr>
<td>a,b,c</td>
</tr>
<tr>
<td>a,c</td>
</tr>
<tr>
<td>a,d</td>
</tr>
<tr>
<td>a,e</td>
</tr>
<tr>
<td>b,c</td>
</tr>
<tr>
<td>b,d</td>
</tr>
<tr>
<td>b,e</td>
</tr>
<tr>
<td>c,d</td>
</tr>
<tr>
<td>c,e</td>
</tr>
</tbody>
</table>

### Itemsets

<table>
<thead>
<tr>
<th>Itemset</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>?</td>
</tr>
<tr>
<td>b</td>
<td>?</td>
</tr>
<tr>
<td>c</td>
<td>?</td>
</tr>
<tr>
<td>d</td>
<td>?</td>
</tr>
<tr>
<td>e</td>
<td>?</td>
</tr>
<tr>
<td>a,b</td>
<td>?</td>
</tr>
<tr>
<td>a,c</td>
<td>?</td>
</tr>
<tr>
<td>a,d</td>
<td>?</td>
</tr>
<tr>
<td>a,e</td>
<td>?</td>
</tr>
<tr>
<td>b,c</td>
<td>?</td>
</tr>
<tr>
<td>b,d</td>
<td>?</td>
</tr>
<tr>
<td>b,e</td>
<td>?</td>
</tr>
<tr>
<td>c,d</td>
<td>?</td>
</tr>
<tr>
<td>c,e</td>
<td>?</td>
</tr>
</tbody>
</table>

### Association Rules

<table>
<thead>
<tr>
<th>Itemset</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>d,e</td>
<td>?</td>
</tr>
<tr>
<td>a,b,c</td>
<td>?</td>
</tr>
<tr>
<td>a,b,d</td>
<td>?</td>
</tr>
<tr>
<td>a,b,e</td>
<td>?</td>
</tr>
<tr>
<td>b,c,d</td>
<td>?</td>
</tr>
<tr>
<td>b,c,e</td>
<td>?</td>
</tr>
<tr>
<td>c,d,e</td>
<td>?</td>
</tr>
<tr>
<td>a,b,c,d</td>
<td>?</td>
</tr>
<tr>
<td>a,b,c,e</td>
<td>?</td>
</tr>
<tr>
<td>a,b,d,e</td>
<td>?</td>
</tr>
<tr>
<td>a,c,d,e</td>
<td>?</td>
</tr>
<tr>
<td>b,c,d,e</td>
<td>?</td>
</tr>
<tr>
<td>a,b,c,d,e</td>
<td>?</td>
</tr>
</tbody>
</table>
## Example

**Data** \( D \)

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>a, b, e</td>
<td>7</td>
</tr>
<tr>
<td>b, c</td>
<td>7</td>
</tr>
<tr>
<td>a, c, e</td>
<td>6</td>
</tr>
<tr>
<td>a, b, c, e</td>
<td>5</td>
</tr>
<tr>
<td>a, b, d, e</td>
<td>5</td>
</tr>
<tr>
<td>b, c, d</td>
<td>2</td>
</tr>
<tr>
<td>a, b, c</td>
<td>4</td>
</tr>
<tr>
<td>a, c</td>
<td>4</td>
</tr>
<tr>
<td>a, d</td>
<td>1</td>
</tr>
<tr>
<td>a, b</td>
<td>5</td>
</tr>
<tr>
<td>a, e</td>
<td>5</td>
</tr>
<tr>
<td>b, c</td>
<td>4</td>
</tr>
<tr>
<td>b, d</td>
<td>2</td>
</tr>
<tr>
<td>b, e</td>
<td>4</td>
</tr>
<tr>
<td>c, d</td>
<td>2</td>
</tr>
<tr>
<td>c, e</td>
<td>2</td>
</tr>
</tbody>
</table>

**Itemsets**

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>d, e</td>
<td>1</td>
</tr>
<tr>
<td>a, b, c</td>
<td>2</td>
</tr>
<tr>
<td>a, b, d</td>
<td>1</td>
</tr>
<tr>
<td>a, b, e</td>
<td>4</td>
</tr>
<tr>
<td>b, c, d</td>
<td>1</td>
</tr>
<tr>
<td>b, c, e</td>
<td>1</td>
</tr>
<tr>
<td>c, d, e</td>
<td>0</td>
</tr>
<tr>
<td>a, b, c, d</td>
<td>0</td>
</tr>
<tr>
<td>a, b, c, e</td>
<td>1</td>
</tr>
<tr>
<td>a, b, d, e</td>
<td>1</td>
</tr>
<tr>
<td>a, c, d, e</td>
<td>0</td>
</tr>
<tr>
<td>b, c, d, e</td>
<td>0</td>
</tr>
<tr>
<td>a, b, c, d, e</td>
<td>0</td>
</tr>
</tbody>
</table>

**Association Rules**

**Summary**
## Example

**Data D**

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>7</td>
</tr>
<tr>
<td>b</td>
<td>7</td>
</tr>
<tr>
<td>c</td>
<td>6</td>
</tr>
<tr>
<td>d</td>
<td>2</td>
</tr>
<tr>
<td>e</td>
<td>5</td>
</tr>
<tr>
<td>a,b</td>
<td>5</td>
</tr>
<tr>
<td>a,c</td>
<td>4</td>
</tr>
<tr>
<td>a,d</td>
<td>1</td>
</tr>
<tr>
<td>a,e</td>
<td>5</td>
</tr>
<tr>
<td>b,c</td>
<td>4</td>
</tr>
<tr>
<td>b,d</td>
<td>2</td>
</tr>
<tr>
<td>b,e</td>
<td>4</td>
</tr>
<tr>
<td>c,d</td>
<td>1</td>
</tr>
<tr>
<td>c,e</td>
<td>2</td>
</tr>
</tbody>
</table>

**Summary**

\[
L = \{\{a\}, \{b\}, \{c\}, \{e\}, \{a, b\}, \{a, c\}, \{a, e\}, \{b, c\}, \{b, e\}, \{a, b, e\}\}
\]

**Itemsets**

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>d,e</td>
<td>1</td>
</tr>
<tr>
<td>a,b,c</td>
<td>2</td>
</tr>
<tr>
<td>a,b,d</td>
<td>1</td>
</tr>
<tr>
<td>a,b,e</td>
<td>4</td>
</tr>
<tr>
<td>b,c,d</td>
<td>1</td>
</tr>
<tr>
<td>b,c,e</td>
<td>1</td>
</tr>
<tr>
<td>c,d,e</td>
<td>0</td>
</tr>
<tr>
<td>a,b,c,d</td>
<td>0</td>
</tr>
<tr>
<td>a,b,c,e</td>
<td>1</td>
</tr>
<tr>
<td>a,b,d,e</td>
<td>1</td>
</tr>
<tr>
<td>a,c,d,e</td>
<td>0</td>
</tr>
<tr>
<td>b,c,d,e</td>
<td>0</td>
</tr>
<tr>
<td>a,b,c,d,e</td>
<td>0</td>
</tr>
</tbody>
</table>
Properties of Naive Algorithm

- returns correct result
- always terminates

But: counting support for each itemset \( X \subseteq \mathcal{P}(I) \) is not applicable as \( |\mathcal{P}(I)| \) is exponential in \( |I| \)
Observations

\[ \text{support}_D(X) \geq \text{support}_D(X \cup Y) \]

- \[ \text{support}_D(X) \geq \theta_s \Rightarrow \forall Y : Y \subset X : \text{support}_D(Y) \geq \theta_s \]
  "all subsets of a frequent set are frequent"

- \[ \text{support}_D(X) < \theta_s \Rightarrow \forall Y : Y \supset X : \text{support}_D(Y) < \theta_s \]
  "all supersets of an infrequent set \( X \) are not frequent"

- example: \[ \text{support}_D(\{a, b\}) \geq \text{support}_D(\{a, b, c, d\}) \]
Apriori Algorithm

- **Breadth-first/ levelwise search**
  1. find frequent itemsets of length 1
  2. find frequent itemsets of length 2
  3. 

- only explores itemsets where all subsets are known to be frequent
Apriori Algorithm

function $\text{APRIORI}(D, \theta_s)$

$k \leftarrow 1$
$L_k \leftarrow \{\{i\} \mid i \in I, \text{support}_D(\{i\}) \geq \theta_s\}$

while $L_k \neq \emptyset$

$C_{k+1} \leftarrow \text{generateCandidates}(L_k, k + 1)$
$L_{k+1} \leftarrow \{X \in C_{k+1} \mid \text{support}_D(X) \geq \theta_s\}$

$k \leftarrow k + 1$

end while

return $\bigcup_{k=1}^{\infty} L_k$

end function
Candidate Generation

generates candidates of length $k$ from frequent itemsets $L$ of length $k - 1$

function $\text{generateCandidates}(L, k)$

$C \leftarrow \{X \cup Y | X, Y \in L \land |X \cup Y| = k\}$

$C \leftarrow \{X \in C | \forall Y \subset X : |Y| = k - 1 \Rightarrow Y \in L\}$

return $C$

end function
Example

Data $D$

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a,b,e</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b,c</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a,c,e</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a,b,c,e</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a,b,d,e</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b,c,d</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a,b,c</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a,c</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a,b,e</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

find itemsets with $\theta_s \geq 0.3$

$X$ frequent $\iff \#_D(X) > 2$
Example

Data $D$

| a, b, e |
| a, c, e |
| a, b, c, e |
| a, b, d, e |
| b, c, d |
| a, b, c |
| a, c |
| a, b, e |

$C_1$  
| a  |
| b  |
| c  |
| d  |
| e  |
# Example

## Data $D$

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>a, b, e</td>
<td>7</td>
</tr>
<tr>
<td>b, c</td>
<td>7</td>
</tr>
<tr>
<td>a, c, e</td>
<td>6</td>
</tr>
<tr>
<td>a, b, c, e</td>
<td>7</td>
</tr>
<tr>
<td>a, b, d, e</td>
<td>6</td>
</tr>
<tr>
<td>b, c, d</td>
<td>2</td>
</tr>
<tr>
<td>a, b, c</td>
<td>5</td>
</tr>
<tr>
<td>a, c</td>
<td>5</td>
</tr>
<tr>
<td>a, b, e</td>
<td>2</td>
</tr>
</tbody>
</table>
Example

**Data** $D$

<table>
<thead>
<tr>
<th>Items</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>a, b, e</td>
<td>7</td>
</tr>
<tr>
<td>b, c</td>
<td>7</td>
</tr>
<tr>
<td>a, c, e</td>
<td>6</td>
</tr>
<tr>
<td>a, b, c, e</td>
<td>5</td>
</tr>
<tr>
<td>a, b, d, e</td>
<td>2</td>
</tr>
<tr>
<td>b, c, d</td>
<td></td>
</tr>
<tr>
<td>a, b, c</td>
<td></td>
</tr>
<tr>
<td>a, c</td>
<td></td>
</tr>
<tr>
<td>a, b, e</td>
<td></td>
</tr>
</tbody>
</table>

### Itemsets

- a, b, e
- b, c
- a, c, e
- a, b, c, e
- a, b, d, e
- b, c, d
- a, b, c
- a, c
- a, b, e

### Association Rules

- a, b, e → c
- b, c → a, c, e
- a, b, c, e → a, b, d, e
- b, c, d → a, b, c
- a, b, c → a, c
- a, b, e → a, b, c
Example

Data $D$

<table>
<thead>
<tr>
<th>Items</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>a, b, e</td>
<td>7</td>
</tr>
<tr>
<td>b, c</td>
<td>7</td>
</tr>
<tr>
<td>a, c, e</td>
<td>6</td>
</tr>
<tr>
<td>a, b, c, e</td>
<td>2</td>
</tr>
<tr>
<td>b, c, d</td>
<td>2</td>
</tr>
<tr>
<td>a, b, c</td>
<td>5</td>
</tr>
<tr>
<td>a, c</td>
<td></td>
</tr>
<tr>
<td>a, b, e</td>
<td></td>
</tr>
</tbody>
</table>

$C_1$ and $L_1$

<table>
<thead>
<tr>
<th>Item</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>7</td>
</tr>
<tr>
<td>b</td>
<td>7</td>
</tr>
<tr>
<td>c</td>
<td>6</td>
</tr>
<tr>
<td>d</td>
<td>2</td>
</tr>
<tr>
<td>e</td>
<td>5</td>
</tr>
</tbody>
</table>

Steffen Rendle Information Systems and Machine Learning Lab (ISMLL), University of Hildesheim
Example

Data $D$

| a, b, e  | a, b  | a, c  | a, e  |
| a, b, c, e | a, c  | a, c  | a, e  |
| a, b, d, e | a, e  | a, e  | a, e  |
| b, c, d  | b, e  | b, e  | b, e  |
| a, b, c  | c, e  | c, e  | c, e  |
| a, c     |       |       |       |
| a, b, e  |       |       |       |

$C_2$ #

| a, b   | ? |
| a, c   | ? |
| a, e   | ? |
| b, c   | ? |
| b, e   | ? |
| c, e   | ? |

$L_1$

| a  |
| b  |
| c  |
| e  |
Example

Data $D$

<table>
<thead>
<tr>
<th>a, b, e</th>
</tr>
</thead>
<tbody>
<tr>
<td>b, c</td>
</tr>
<tr>
<td>a, c, e</td>
</tr>
<tr>
<td>a, b, c, e</td>
</tr>
<tr>
<td>a, b, d, e</td>
</tr>
<tr>
<td>b, c, d</td>
</tr>
<tr>
<td>a, b, c</td>
</tr>
<tr>
<td>a, c</td>
</tr>
<tr>
<td>a, b, e</td>
</tr>
</tbody>
</table>

$C_2$   | #  |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a, b</td>
<td>5</td>
</tr>
<tr>
<td>a, c</td>
<td>4</td>
</tr>
<tr>
<td>a, e</td>
<td>5</td>
</tr>
<tr>
<td>b, c</td>
<td>4</td>
</tr>
<tr>
<td>b, e</td>
<td>4</td>
</tr>
<tr>
<td>c, e</td>
<td>2</td>
</tr>
</tbody>
</table>
Example

<table>
<thead>
<tr>
<th>Data D</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a, b, e</td>
<td></td>
</tr>
<tr>
<td>b, c</td>
<td></td>
</tr>
<tr>
<td>a, c, e</td>
<td></td>
</tr>
<tr>
<td>a, b, c, e</td>
<td></td>
</tr>
<tr>
<td>a, b, d, e</td>
<td></td>
</tr>
<tr>
<td>b, c, d</td>
<td></td>
</tr>
<tr>
<td>a, b, c</td>
<td></td>
</tr>
<tr>
<td>a, c</td>
<td></td>
</tr>
<tr>
<td>a, b, e</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C₂</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>a, b</td>
<td>5</td>
</tr>
<tr>
<td>a, c</td>
<td>4</td>
</tr>
<tr>
<td>a, e</td>
<td>5</td>
</tr>
<tr>
<td>b, c</td>
<td>4</td>
</tr>
<tr>
<td>b, e</td>
<td>4</td>
</tr>
<tr>
<td>c, e</td>
<td>2</td>
</tr>
</tbody>
</table>
Example

Data $D$

<table>
<thead>
<tr>
<th>a,b,e</th>
<th>b,c</th>
<th>a,c,e</th>
<th>a,b,c,e</th>
<th>a,b,d,e</th>
<th>b,c,d</th>
<th>a,b,c</th>
<th>a,c</th>
<th>a,b,e</th>
</tr>
</thead>
</table>

C2

<table>
<thead>
<tr>
<th>a,b</th>
<th>a,c</th>
<th>a,e</th>
<th>b,c</th>
<th>b,e</th>
<th>c,e</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

L2

<table>
<thead>
<tr>
<th>a,b</th>
<th>a,c</th>
<th>a,e</th>
<th>b,c</th>
<th>b,e</th>
</tr>
</thead>
</table>

Steffen Rendle
Information Systems and Machine Learning Lab (ISMLL), University of Hildesheim
Example

### Data $D$

<table>
<thead>
<tr>
<th>Itemset</th>
</tr>
</thead>
<tbody>
<tr>
<td>a,b,e</td>
</tr>
<tr>
<td>b,c</td>
</tr>
<tr>
<td>a,c,e</td>
</tr>
<tr>
<td>a,b,c,e</td>
</tr>
<tr>
<td>a,b,d,e</td>
</tr>
<tr>
<td>b,c,d</td>
</tr>
<tr>
<td>a,b,c</td>
</tr>
<tr>
<td>a,c</td>
</tr>
<tr>
<td>a,b,e</td>
</tr>
</tbody>
</table>

### Frequent Itemsets $C_3$

<table>
<thead>
<tr>
<th>Itemset</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>a,b,c</td>
<td>?</td>
</tr>
<tr>
<td>a,b,e</td>
<td>?</td>
</tr>
<tr>
<td>a,c,e</td>
<td>?</td>
</tr>
<tr>
<td>b,c,e</td>
<td>?</td>
</tr>
</tbody>
</table>

### Association Rules $L_2$

<table>
<thead>
<tr>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>a,b</td>
</tr>
<tr>
<td>a,c</td>
</tr>
<tr>
<td>a,e</td>
</tr>
<tr>
<td>b,c</td>
</tr>
<tr>
<td>b,e</td>
</tr>
</tbody>
</table>
Example

Data $D$

<table>
<thead>
<tr>
<th>a, b, e</th>
</tr>
</thead>
<tbody>
<tr>
<td>b, c</td>
</tr>
<tr>
<td>a, c, e</td>
</tr>
<tr>
<td>a, b, c, e</td>
</tr>
<tr>
<td>a, b, d, e</td>
</tr>
<tr>
<td>b, c, d</td>
</tr>
<tr>
<td>a, b, c</td>
</tr>
<tr>
<td>a, c</td>
</tr>
<tr>
<td>a, b, e</td>
</tr>
</tbody>
</table>

$C_3$ | # |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a, b, c</td>
<td>?</td>
</tr>
<tr>
<td>a, b, e</td>
<td>?</td>
</tr>
</tbody>
</table>

Pruning: \{c, e\} $\not\subseteq L_2$
# Example

**Data $D$**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a,b,e</td>
<td></td>
</tr>
<tr>
<td>b,c</td>
<td></td>
</tr>
<tr>
<td>a,c,e</td>
<td></td>
</tr>
<tr>
<td>a,b,c,e</td>
<td></td>
</tr>
<tr>
<td>a,b,d,e</td>
<td></td>
</tr>
<tr>
<td>b,c,d</td>
<td></td>
</tr>
<tr>
<td>a,b,c</td>
<td></td>
</tr>
<tr>
<td>a,c</td>
<td></td>
</tr>
<tr>
<td>a,b,e</td>
<td></td>
</tr>
</tbody>
</table>

**$C_3$**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a,b,c</td>
<td>2</td>
</tr>
<tr>
<td>a,b,e</td>
<td>4</td>
</tr>
</tbody>
</table>
### Example

<table>
<thead>
<tr>
<th>Data $D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>a,b,e</td>
</tr>
<tr>
<td>b,c</td>
</tr>
<tr>
<td>a,c,e</td>
</tr>
<tr>
<td>a,b,c,e</td>
</tr>
<tr>
<td>a,b,d,e</td>
</tr>
<tr>
<td>b,c,d</td>
</tr>
<tr>
<td>a,b,c</td>
</tr>
<tr>
<td>a,c</td>
</tr>
<tr>
<td>a,b,e</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$C_3$</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>a,b,c</td>
<td>2</td>
</tr>
<tr>
<td>a,b,e</td>
<td>4</td>
</tr>
</tbody>
</table>
Example

Data $D$

| a,b,e  | b,c  | a,c,e  | a,b,c,e  | a,b,d,e  | b,c,d  | a,b,c  | a,c  | a,b,e  |

<table>
<thead>
<tr>
<th>$C_3$</th>
<th>$#$</th>
</tr>
</thead>
<tbody>
<tr>
<td>a,b,c</td>
<td>2</td>
</tr>
<tr>
<td>a,b,e</td>
<td>4</td>
</tr>
</tbody>
</table>

$C_3 \rightarrow a,b,c$  
$L_3 \rightarrow a,b,e$
**Example**

<table>
<thead>
<tr>
<th>Data $D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>a,b,e</td>
</tr>
<tr>
<td>b,c</td>
</tr>
<tr>
<td>a,c,e</td>
</tr>
<tr>
<td>a,b,c,e</td>
</tr>
<tr>
<td>a,b,d,e</td>
</tr>
<tr>
<td>b,c,d</td>
</tr>
<tr>
<td>a,b,c</td>
</tr>
<tr>
<td>a,c</td>
</tr>
<tr>
<td>a,b,e</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$C_4$</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$L_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>a,b,e</td>
</tr>
</tbody>
</table>

Steffen Rendle
Information Systems and Machine Learning Lab (ISMLL), University of Hildesheim
Example

Data $D$

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a, b, e</td>
<td>b, c</td>
<td>a, c, e</td>
</tr>
<tr>
<td>a, b, c, e</td>
<td>a, b, d, e</td>
<td>a, b, c</td>
</tr>
<tr>
<td>b, c, d</td>
<td>a, b, c</td>
<td>a, c</td>
</tr>
<tr>
<td>a, b, c</td>
<td>a, c</td>
<td>a, b, e</td>
</tr>
<tr>
<td>a, c</td>
<td>b, e</td>
<td>b, e</td>
</tr>
<tr>
<td>b, e</td>
<td>a, b, e</td>
<td>L_1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>L_2</td>
</tr>
<tr>
<td>a, b</td>
<td>a, c</td>
<td>a, e</td>
</tr>
<tr>
<td>a, c</td>
<td>b, c</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>L_3</td>
</tr>
<tr>
<td>a, b, e</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$L = \{\{a\}, \{b\}, \{c\}, \{e\}, \{a, b\}, \{a, c\}, \{a, e\}, \{b, c\}, \{b, e\}, \{a, b, e\}\}$
Trie / Prefix Tree

For candidate generation and frequency counting, a trie can be used:

- a trie is a tree
- each node contains an item and a frequency counter
- each path from the root to a node corresponds to an itemset
- the k-th level represents itemsets of length $k$
- the items in a trie are ordered
Example

Data $D$

- a, b, e
- b, c
- a, c, e
- a, b, c, e
- a, b, d, e
- b, c, d
- a, b, c
- a, c
- a, b, e

Prefix Tree

- $\emptyset$
- a 7
- b 7
- c 6
- e 5

- b 5
- c 4
- e 5
- c 4
- e 4
- e 4

Steffen Rendle
Information Systems and Machine Learning Lab (ISMLL), University of Hildesheim
Example

Data $D$

<table>
<thead>
<tr>
<th>Itemsets</th>
<th>Association Rules</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>a,b,e</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b,c</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a,c,e</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a,b,c,e</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a,b,d,e</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b,c,d</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a,b,c</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a,c</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a,b,e</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$\#\{a,b,e\} = 4$

Steffen Rendle

Information Systems and Machine Learning Lab (ISMLL), University of Hildesheim
Example

Data $D$

| a,b,e |
| b,c |
| a,c,e |
| a,b,c,e |
| a,b,d,e |
| b,c,d |
| a,b,c |
| a,c |
| a,b,e |

$\#\{b\} = 7$
Example

Data $D$

- a, b, e
- b, c
- a, c, e
- a, b, c, e
- a, b, d, e
- b, c, d
- a, b, c
- a, c
- a, b, e

$\{b, e\} = 4$

# Example

<table>
<thead>
<tr>
<th>Pattern Mining</th>
<th>Itemsets</th>
<th>Association Rules</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Steffen Rendle</td>
<td>Information Systems and Machine Learning Lab (ISMLL), University of Hildesheim</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Trie: Frequency Counting

To count frequencies with a trie, each transaction $d \in D$ is handled the following way:

1. sort $d$
2. start at the root
3. for each item $i \in d$ follow the node $i$, increase it by one and recursively repeat this for $d \setminus \{i\}$
Trie: Candidate Generation

Candidates of length $k$ can be generated from a trie of depth $k - 1$:

1. for each node at level $k - 1$ append its siblings
2. prune infrequent childs
Trie: Example

<table>
<thead>
<tr>
<th>Data $D$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a, b, e</td>
<td></td>
</tr>
<tr>
<td>b, c</td>
<td></td>
</tr>
<tr>
<td>a, c, e</td>
<td></td>
</tr>
<tr>
<td>a, b, c, e</td>
<td></td>
</tr>
<tr>
<td>a, b, d, e</td>
<td></td>
</tr>
<tr>
<td>b, c, d</td>
<td></td>
</tr>
<tr>
<td>a, b, c</td>
<td></td>
</tr>
<tr>
<td>a, c</td>
<td></td>
</tr>
<tr>
<td>a, b, e</td>
<td></td>
</tr>
</tbody>
</table>

Find itemsets with $\theta_s \geq 0.3$

$X$ frequent $\iff \#_D(X) > 2$

... see blackboard ...
Eclat Algorithm

- Algorithm for itemset mining
- Depth-first algorithm
- Vertical data base layout
  - For each pattern: store the cover, i.e. all transactions that include this pattern. e.g. \((a, \{d_1, d_3, d_4, d_5, d_7, d_8, d_9\})\)
  - Count frequency by intersection
Eclat Algorithm

\textbf{function} \texttt{ECLAT}(D, \theta_s) \\
\hspace{1em} C_\emptyset = \{(i, \{d \in D \mid i \in d\}) \mid i \in I\} \\
\hspace{1em} L_\emptyset = \left\{(i, D_i) \in C_\emptyset \mid \frac{|D_i|}{|D|} \geq \theta_s \right\} \\
\textbf{return} \texttt{ECLATRECURSION}(L_\emptyset, \emptyset, \theta_s) \\
\textbf{end function}
Eclat Algorithm

function ECLATRECURSION(L, p, θs)
    F ← ∅
    for all (i, Di) ∈ L do
        q ← p ∪ {i}
        F ← F ∪ {p}
        Cq ← {(j, Di ∩ Dj) | (j, Dj) ∈ L, j > i}
        Lq ← \{ (k, Dk) ∈ Cq | \frac{|D_k|}{|D|} \geq θ_s \}
        if Lq ≠ ∅ then
            F ← F ∪ ECLATRECURSION(Lq, q, θs)
        end if
    end for
    return F
end function
Example

<table>
<thead>
<tr>
<th>Data $D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>a, b, e</td>
</tr>
<tr>
<td>b, c</td>
</tr>
<tr>
<td>a, c, e</td>
</tr>
<tr>
<td>a, b, c, e</td>
</tr>
<tr>
<td>a, b, d, e</td>
</tr>
<tr>
<td>b, c, d</td>
</tr>
<tr>
<td>a, b, c</td>
</tr>
<tr>
<td>a, c</td>
</tr>
<tr>
<td>a, b, e</td>
</tr>
</tbody>
</table>

find itemsets with $\theta_s \geq 0.3$

$X$ frequent $\iff \#_D(X) > 2$

... see blackboard ...
Association Rules

- Which itemsets $Y$ occur often if another itemset $X$ appears? $X \Rightarrow Y$
- Example: a customer buying diapers also buys beer
  \{diapers\} $\Rightarrow$ \{beer\}
- Given
  - Items $I = \{i_1, \ldots, i_m\}$
  - Data $D \subseteq \mathcal{P}(I)$ multiset
  - Frequency thresholds $\theta_s$
  - Confidence threshold $\theta_c$
- To find
  - Rules $R = \{X \Rightarrow Y | support_D(X \Rightarrow Y) \geq \theta_s \land confidence_D(X \Rightarrow Y) \geq \theta_c\}$
Definitions and Terms

▶ **support** measures how often the rule \( X \Rightarrow Y \) appears

- \( \text{support}_D(X \Rightarrow Y) = \text{support}_D(X \cup Y) \)
- \( X \Rightarrow Y \) is **frequent** / **large** iff \( \text{support}_D(X \Rightarrow Y) \geq \theta_s \)

▶ **confidence** measures how likely it is that \( Y \) appears if \( X \) is present.

- \( \text{confidence}_D(X \Rightarrow Y) = \frac{\text{support}_D(X \Rightarrow Y)}{\text{support}_D(X)} \)

▶ for a rule \( X \Rightarrow Y \), \( Y \) is called **head** and \( X \) is called **body**
Example

Data $D$

| a, b, e | b, c | a, c, e | a, b, c, e | a, b, d, e | b, c, d | a, b, c | a, c | a, b, e |

find rules with $\theta_s \geq 0.3$

$X$ frequent $\iff \#_D(X) > 2$

find rules with $\theta_c \geq 0.8$
Example

<table>
<thead>
<tr>
<th>Data $D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>a,b,e</td>
</tr>
<tr>
<td>b,c</td>
</tr>
<tr>
<td>a,c,e</td>
</tr>
<tr>
<td>a,b,c,e</td>
</tr>
<tr>
<td>a,b,d,e</td>
</tr>
<tr>
<td>b,c,d</td>
</tr>
<tr>
<td>a,b,c</td>
</tr>
<tr>
<td>a,c</td>
</tr>
<tr>
<td>a,b,e</td>
</tr>
</tbody>
</table>

$Confidence_D(X \Rightarrow Y) = \frac{support_D(X \Rightarrow Y)}{support_D(X)}$

$support_D(X \Rightarrow Y) = support_D(X \cup Y)$

$L = \{\{a\}, \{b\}, \{c\}, \{e\}, \{a, b\}, \{a, c\}, \{a, e\}, \{b, c\}, \{b, e\}, \{a, b, e\}\}$

\ldots see blackboard \ldots
Example

Data $D$

<table>
<thead>
<tr>
<th>a,b,e</th>
</tr>
</thead>
<tbody>
<tr>
<td>b,c</td>
</tr>
<tr>
<td>a,c,e</td>
</tr>
<tr>
<td>a,b,c,e</td>
</tr>
<tr>
<td>a,b,d,e</td>
</tr>
<tr>
<td>b,c,d</td>
</tr>
<tr>
<td>a,b,c</td>
</tr>
<tr>
<td>a,c</td>
</tr>
<tr>
<td>a,b,e</td>
</tr>
</tbody>
</table>

$R = \{e \Rightarrow a, e \Rightarrow b, e \Rightarrow ab, ab \Rightarrow e, ae \Rightarrow b, be \Rightarrow a\} \cup \{X \Rightarrow \emptyset | X \in L\}$
Observations

▶ expanding the head of a rule by an item of the body, results in a rule with less or equal confidence.

\[
\text{confidence}_D(X \setminus Z \Rightarrow Y \cup Z) \leq \text{confidence}_D(X \Rightarrow Y)
\]

▶ proof:

\[
\text{confidence}_D(X \setminus Z \Rightarrow Y \cup Z) = \frac{\text{support}_D((X \setminus Z) \cup (Y \cup Z))}{\text{support}_D(X \setminus Z)} \leq \frac{\text{support}_D(X \cup Y)}{\text{support}_D(X)} = \text{confidence}_D(X \Rightarrow Y)
\]

▶ example:

\[
\text{confidence}_D(\{a, b\} \Rightarrow \{c, d\}) \leq \text{confidence}_D(\{a, b, c\} \Rightarrow \{d\})
\]
Algorithm

Association rule mining is done in two steps:
1. find frequent itemsets (see itemset mining)
2. extract rules from the frequent itemsets
AssociationRules Algorithm

\textbf{function} \textsc{AssociationRules}(D, \theta_s, \theta_c)

\begin{align*}
L & \leftarrow \text{Apriori}(D, \theta_s) \\
R & \leftarrow \emptyset \\
\text{for all } I \in L \text { do} \\
& \quad k \leftarrow 1 \\
& \quad C_k \leftarrow \{\{i\} | i \in I\} \\
\text{while } C_k \neq \emptyset \text { do} \\
& \quad H_k \leftarrow \{X \in C_k| \text{confidence}_D(I \setminus X \Rightarrow X) \geq \theta_c\} \\
& \quad C_{k+1} \leftarrow \text{generateCandidateHeads}(H_k, k + 1) \\
& \quad k \leftarrow k + 1 \\
\text{end while} \\
R & \leftarrow R \cup \{I \setminus X \Rightarrow X | X \in \bigcup_{k=1}^{\infty} H_k\} \cup \{I \Rightarrow \emptyset\}
\end{align*}

\textbf{end for}

\textbf{return } R

\textbf{end function}
Candidate Generation for Heads of Rules

generates candidate heads of length $k$ from heads $H$ of length $k - 1$

```plaintext
function GENERATE_CANDIDATE_HEADS(H, k)
    C ← \{X ∪ Y | X, Y ∈ H ∧ |X ∪ Y| = k\}
    C ← \{X ∈ C | ∀ Y ⊂ X : |Y| = k - 1 ⇒ Y ∈ H\}
    return C
end function
```
Remarks

- Calculating the confidence can be reduced to calculating the support:

\[
\text{confidence}_D(I \setminus X \Rightarrow X) = \frac{\text{support}_D((I \setminus X) \cup X)}{\text{support}_D(I \setminus X)} \geq \theta_c
\]

\[
\iff \frac{\text{support}_D(I)}{\text{support}_D(I \setminus X)} \geq \theta_c
\]

\[
\iff \text{support}_D(I \setminus X) \leq \frac{1}{\theta_c} \text{support}_D(I)
\]

- If \( \theta_c \geq \theta_s \), the values for \( \text{support}_D \) can be looked up in the trie and no database pass is necessary.
Example

Trace of inner loop of the algorithm `ASSOCIATIONRULES` for
\( I = \{a, b, e\} \).

\[ \ldots \text{see blackboard} \ldots \]
Outlook

- Extensions to Apriori and Eclat
- Further pattern: sequences, trees, ...
- Background knowledge: e.g. taxonomies
Sequence mining

Takes the time into account, when an action is performed. E.g.

- A database of courses attended by a student in one term, i.e. sequences of sets:
  - Student1: (\{linear algebra, c++, algorithm theory\}, \{machine learning, numerics, economics\}, \{bayesian networks\})
  - Student2: (\{linear algebra, java\}, \{software engineering\}, \{numerics\})
  - Student3: (\{linear algebra, java, algorithm theory\}, \{economics\}, \{machine learning, numerics\}, \{bayesian networks\})
  - ...

- A frequent sequence might be (\{linear algebra, algorithm theory\}, \{machine learning, numerics\}, \{bayesian networks\})
Use taxonomies

Background knowledge in terms of taxonomies might be used for mining patterns. E.g.

- The following taxonomy is given over subjects
  - linear algebra isa mathematics
  - mathematics isa science
  - computer science isa science

- In the student database one could mine the association rule using the taxonomy:
  
  *if someone has attended machine learning then (s)he also has attended some mathematic lecture*

  \{machine learning\} $\Rightarrow$ \{mathematics\}
Conclusion

- Task: Finding frequent patterns in database.
- Efficient algorithms explore only promising candidates by pruning.
- Mining association rules can be reduced to mining itemsets with an additional post processing step.
R. Agrawal and R. Srikant.
Fast algorithms for mining association rules.

R. Agrawal and R. Srikant.
Mining sequential patterns.

L. Schmidt-Thieme.
Algorithmic features of eclat.
2004.