Exercise 4: Sequential Patterns + Process Mining + kmeans

Knowledge Discovery in Databases I
SS 2016
Sequence Database
<(ABBA)(BBACA)(CBAA)
(ACA)(BAAB)>

Ex4.1: Sequential Pattern Mining

\[ D_\emptyset \]

\[ D_A \]

\[ D_B \]

\[ D_C \]
**Ex4.1: Sequential Pattern Mining**

<table>
<thead>
<tr>
<th>( D_A )</th>
<th>( S_1 )</th>
<th>BBA</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S_2 )</td>
<td>CA</td>
<td></td>
</tr>
<tr>
<td>( S_3 )</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>( S_4 )</td>
<td>CA</td>
<td></td>
</tr>
<tr>
<td>( S_5 )</td>
<td>AB</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( A(5), B(2), C(2) )</td>
<td></td>
</tr>
</tbody>
</table>

\[
\begin{align*}
D_{AA} & \\
S_1 & \text{ } \\
S_2 & \text{ } \\
S_3 & \text{ } \\
S_4 & \text{ } \\
S_5 & B \\
\end{align*}
\]

\[
\begin{align*}
D_{AB} & \\
S_1 & BA \\
S_2 & \text{ } \\
S_3 & \text{ } \\
S_4 & \text{ } \\
S_5 & \text{ } \\
\end{align*}
\]

\[
\begin{align*}
D_{AC} & \\
S_1 & \text{ } \\
S_2 & A \\
S_3 & \text{ } \\
S_4 & A \\
S_5 & \text{ } \\
\end{align*}
\]

\[
\begin{align*}
D_{ACA} & \\
S_1 & \text{ } \\
S_2 & \text{ } \\
S_3 & \text{ } \\
S_4 & \text{ } \\
S_5 & A(0) \\
\end{align*}
\]

Knowledge Discovery in Databases I: Exercise 4
Ex4.1: Sequential Pattern Mining

Knowledge Discovery in Databases I: Exercise 4
### Ex4.1: Sequential Pattern Mining

<table>
<thead>
<tr>
<th>$D_C$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td></td>
</tr>
<tr>
<td>$S_2$</td>
<td>$A$</td>
</tr>
<tr>
<td>$S_3$</td>
<td>$BAA$</td>
</tr>
<tr>
<td>$S_4$</td>
<td>$A$</td>
</tr>
<tr>
<td>$S_5$</td>
<td></td>
</tr>
</tbody>
</table>

$A(3), B(1), C(0)$

<table>
<thead>
<tr>
<th>$D_{CA}$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td></td>
</tr>
<tr>
<td>$S_2$</td>
<td></td>
</tr>
<tr>
<td>$S_3$</td>
<td>$A$</td>
</tr>
<tr>
<td>$S_4$</td>
<td></td>
</tr>
<tr>
<td>$S_5$</td>
<td></td>
</tr>
</tbody>
</table>

$A(1)$
Ex4.1: Sequential Pattern Mining

Frequent patterns:

\( A, AA, AB, AC, ACA, B, BA, BAA, BB, BBA, C, CA \)
Frequent patterns:
\(A, AA, AB, AC, ACA, B, BA, BAA, BB, BBA, C, CA\)
Mining for consecutive patterns

1) Large frequent patterns can start later. Consider all starting points when projecting in the first step:

\[ ABAAABCD \]
\[ \Rightarrow AAABCD \land CD \in D_{AB} \]

2) Do not skip objects. Next one is important for later projections!

3) Pruning strategy has to be changed
Log $L = \{ac, abc, abbc, abbbbc\}$

Obviously:
Ex4.2: $\alpha$-Algorithm

\[ \text{Log } L = \{ac, abc, abbc, abbbc\} \]

Due to $A = \{a\}, B = \{c\}$
Log $L = \{ac, abc, abbc, abbbc\}$

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>#</td>
<td>→</td>
<td>→</td>
</tr>
<tr>
<td>b</td>
<td>←</td>
<td>‖</td>
<td>→</td>
</tr>
<tr>
<td>c</td>
<td>←</td>
<td>←</td>
<td>#</td>
</tr>
</tbody>
</table>

Other possibilities for a place $p$ and two sets $A \rightarrow p \rightarrow B$?

$b \notin A, B$ because $b \neq b$

Observation: $\|$ on diagonal (= 1-loop) impossible for $\alpha$–Algorithm
Mapping from sequences to $\mathbb{R}^2$:

„Bag-of-Words-Model“ defines classes (→ vector components) and stores support of these classes in a vector.

In the sequence case, we lose order information.

<table>
<thead>
<tr>
<th>day</th>
<th>sequence</th>
<th>Vector repr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>monday</td>
<td>bbabb</td>
<td>(1,4)</td>
</tr>
<tr>
<td>tuesday</td>
<td>aabaa</td>
<td>(4,1)</td>
</tr>
<tr>
<td>wednesday</td>
<td>ababaaa</td>
<td>(5,2)</td>
</tr>
<tr>
<td>thursday</td>
<td>baabababa</td>
<td>(5,3)</td>
</tr>
<tr>
<td>friday</td>
<td>bbaababa</td>
<td>(4,3)</td>
</tr>
<tr>
<td>saturday</td>
<td>babbbb</td>
<td>(1,5)</td>
</tr>
<tr>
<td>sunday</td>
<td>bbbabbbba</td>
<td>(2,6)</td>
</tr>
</tbody>
</table>
Ex4.3: $k$-means
Ex4.3: $k$-means
Ex4.3: k-means

Euklidische Isodistanz
Ex4.3: \( k \)-means
Ex4.3: $k$-means
Ex4.3: $k$-means
Ex4.3: $k$-means
Ex4.3: *k*-means

done!