Pattern mining in personal demographic trajectories

Dmitry I. Ignatov\textsuperscript{1}, Danil Gizdatullin\textsuperscript{1}, Ekaterina Mitrofanova\textsuperscript{1}, Anna Muratova\textsuperscript{1}, Jaume Baixeries\textsuperscript{2}

\textsuperscript{1}National Research University Higher School of Economics, Moscow
\textsuperscript{2}Universitat Politècnica de Catalunya, Barcelona

2016, Moscow
Possible life events

- First job (job)
- The highest education degree is obtained (education)
- Leaving parents’ home (separation)
- First partner (partner)
- First marriage (marriage)
- First child birth (children)
- Break-up (parting)
- ... (divorce)
Generation and Gender Survey (GGS): three waves panel data for 11 generations of Russian citizens starting from 30s

Binary classification
1545 men
3312 women

Examples of sequential patterns
- \(\langle\{education, separation\}, \{work\}, \{marriage\}, \{children\}\rangle(m)\)
- \(\langle\{work\}, \{marriage\}, \{children\}\{education\}\rangle(f)\)
- \(\langle\{partner\}, \{marriage, separation\}, \{children\}\rangle(f)\)
Basic definitions
Textbooks of Han et al., Zaki & Meira, Aggarwal et al., etc

- $s = \langle s_1, ..., s_k \rangle$ is the **subsequence** of $s' = \langle s'_1, ..., s'_k \rangle$ ($s \preceq s'$) if $k \leq k'$ and there exist $1 \leq r_1 < r_2 < ... < r_k \leq k'$ such $s_j = s'_{r_j}$ for all $1 \leq j \leq k$.

- $\text{support}(s, D)$ is the **support** of a sequence $s$ in $D$, i.e. the number of sequences in $D$ such that $s$ is their subsequence.

$$\text{support}(s, D) = |\{s'| s' \in D, s \preceq s'\}|$$

- $s$ is a **frequent closed sequence (sequential pattern)** if there is no $s'$ such that $s \prec s'$ and

$$\text{support}(s, D) = \text{support}(s', D)$$
Let $D$ be a set of sequences:

<table>
<thead>
<tr>
<th></th>
<th>$s_1$</th>
<th>$s_2$</th>
<th>$s_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D$</td>
<td>${a, b, c}{a, b}{b}$</td>
<td>${a}{a, c}{a}$</td>
<td>${a, b}{b, c}$</td>
</tr>
</tbody>
</table>

- $I = \{a, b, c\}$ is the set of all items (atomic events)
- $\langle\{a, b\}\{b\}\rangle$ belongs to $s_1$ and $s_3$ but it is missing in $s_2$
- $\text{support}_D(\langle\{a, b\}\{b\}\rangle) = 2$
- $\{\langle\{a\}\rangle, \langle\{c\}\rangle, \langle\{a\}\{c\}\rangle, \langle\{a, b\}\{b\}\rangle, \langle\{a, c\}\{a\}\rangle\}$ is the set of closed sequences.
Contiguous prefix-based sequential patterns

- \( s = \langle s_1, \ldots, s_k \rangle \) is a contiguous prefix-based subsequence of \( s' = \langle s'_1, \ldots, s'_k \rangle \) (\( s* = s' \)) if \( k \leq k' \) and \( \forall i \in k' : s_i = s'_i \).

- **Support of contiguous prefix-based sequences**
  Let \( T \) be a set of sequences.

  \[
  \text{support}(s, T) = \frac{|\{s' | s' \in T, s* = s'\}|}{|T|}
  \]
Let $0 < \text{minSup} \leq 1$ be a minimal support parameter and $D$ is a set of sequences then searching for prefix-based contiguous sequential patterns is the task of enumeration of all prefix-based contiguous sequences $s$ such that $\text{support}(s, D) \geq \text{minSup}$. Every sequence $s$ with $\text{support}(s, D) \geq \text{minSup}$ is called a prefix-based contiguous sequential pattern.

Prefix-based contiguous sequential pattern (PGSP) $p$ is called closed if there is no PGSP $d$ of greater or equal support such that $d = p$.
Contiguous sequential patterns

Example

Таблица: $D$ is a set of sequences.

<table>
<thead>
<tr>
<th>$s_1$</th>
<th>${a}{b}{d}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_2$</td>
<td>${a}{b}{c}$</td>
</tr>
<tr>
<td>$s_3$</td>
<td>${a, b}{b, c}$</td>
</tr>
</tbody>
</table>

$s = \langle\{a\}\{b\}\rangle$

- $I = \{a, b, c\}$ is the set of all items (atomic events)
- $s_1 = s^*; s_2 = s^*$
- $s_3 \neq s^*$
- $\text{Supp}_D(s) = \frac{2}{3}$
- $\langle\{a\}\{b\}\rangle$ is closed, $\langle\{a\}\rangle$ is not closed.
Growth Rate

\[
growth_{\text{rate}}_{D' \rightarrow D''}(X) = \begin{cases} 
\frac{\text{supp}_{D''}(X)}{\text{supp}_{D'}(X)} & \text{if } \text{supp}_{D'}(X) \neq 0 \\
0 & \text{if } \text{supp}_{D''}(X) = \text{supp}(X) = 0 \\
\infty & \text{if } \text{supp}_{D''}(X) \neq 0 \text{ and } \text{supp}_{D'}(X) = 0
\end{cases}
\]

Class score

\[
\text{score}(s, C) = \sum_{e \subseteq s, e \in E(c)} \frac{growth_{\text{rate}}_C(e)}{growth_{\text{rate}}_C(e) + 1} \cdot \text{supp}_C(e)
\]
s is a new object

\[ \text{normal}_{\oplus}(s) = \frac{\sum_{p \in P_{\oplus} : p \subseteq s} \text{GrowthRate}(p, K_{\oplus}, K_{\ominus})}{\text{median(\text{GrowthRate}(P_{\oplus}))}} \]

\[ \text{normal}_{\ominus}(s) = \frac{\sum_{p \in P_{\ominus} : p \subseteq s} \text{GrowthRate}(p, K_{\ominus}, K_{\oplus})}{\text{median(\text{GrowthRate}(P_{\ominus}))}} \]

Classification via emerging patterns

\[
\text{class}(s) = \begin{cases} 
\text{positive if } \text{normal}_{\oplus}(s) > \text{normal}_{\ominus}(s) \\
\text{negative if } \text{normal}_{\oplus}(s) < \text{normal}_{\ominus}(s) \\
\text{undetermined if } \text{normal}_{\oplus}(s) = \text{normal}_{\ominus}(s)
\end{cases}
\]
Execution example

Input sequences

class 0 : {⟨{a}{b}{c}⟩, ⟨{b}{a}{c}⟩, ⟨{b}{a}{c}⟩, ⟨{b}{c}⟩}
class 1 : {⟨{a}{c}{b}⟩, ⟨{b}{c}{a}⟩, ⟨{b}{c}{a}⟩}

Prefix tree

```
    ∅
   / \
  a(1;1) b(2;2)
 /     /     \
 b(1;0) c(0;1) a(2;0) c(1;2)
 /     /     \
 c(1;0) b(0;1) c(2;0) a(0;2)
```
Counting Growth Rate

\[ \emptyset \]

- \( a(0.75; 1.33) \)
  - \( b(\infty; 0) \)
  - \( c(0; \infty) \)
- \( b(\infty; 0) \)
- \( c(0; \infty) \)
- \( a(\infty; 0) \)
  - \( c(0.38; 2.67) \)
  - \( a(0; \infty) \)

Growth rate

\[
0.75 = \frac{1}{4} / \frac{1}{3} ; 1.33 = \frac{1}{3} / \frac{1}{4} \\
0.38 = \frac{1}{4} / \frac{2}{3} ; 2.67 = \frac{2}{3} / \frac{1}{4}
\]
Computing Score

\[
\emptyset
\]

\[
\begin{array}{c}
\text{a}(0.75; 1.33) \\
\text{b}(\infty; 0) \\
\text{c}(\infty; 0)
\end{array}
\quad
\begin{array}{c}
\text{b}(0.75; 1.33) \\
\text{c}(0; \infty) \\
\text{a}(\infty; 0) \\
\text{c}(0.38; 2.67)
\end{array}
\]

New sequence

\[
\text{minGR} = 2
\]

\[
\langle \{b\}; \{c\}; \{a\} \rangle - ???
\]

\[
\text{Score}_0 = 0
\]

\[
\text{Score}_1 = 2.67 + \infty = \infty
\]
Comparison of closed and non-closed patterns

Рис.: TPR vs FPR for closed and non-closed patterns
Experiments and results

Рис.: TPR-FPR for classification by gender via contiguous prefix-based patterns
Interesting patterns (women)

\[
\langle\{\text{work, separation}\}, \{\text{marriage}\}, \{\text{children}\}, \{\text{education}\}\rangle, [\infty, 0.006]
\]

\[
\langle\{\text{separation, partner}\}, \{\text{marriage}\}\rangle, [\infty, 0.006]
\]

\[
\langle\{\text{work, separation}\}, \{\text{marriage}\}, \{\text{children}\}\rangle, [\infty, 0.008]
\]

\[
\langle\{\text{work, separation}\}, \{\text{marriage}\}\rangle, [\infty, 0.009]
\]
Interesting patterns (men)

\((\langle\{\text{education}\}, \{\text{marriage}\}, \{\text{work}\}, \{\text{children}\}, \{\text{separation}\}\rangle, [10.6, 0.006])\)

\((\langle\{\text{education}\}, \{\text{marriage}\}, \{\text{work}\}, \{\text{children}\}\rangle, [12.7, 0.007])\)

\((\langle\{\text{educ}\}, \{\text{work}\}, \{\text{part}\}, \{\text{mar}\}, \{\text{sep}\}, \{\text{ch}\}\rangle, [10.6, 0.006])\)
Experiments and results

Рис.: TPR-FPR for classification by generation via contiguous prefix-based patterns
Interesting patterns (Different Generations; Women)

Old women

- $\langle\{\text{work}\}, \{\text{separation}\}\rangle$, $[1.85, 0.38]$
- $\langle\{\text{work}\}, \{\text{marriage}, \text{separation}\}\rangle$, $[3.92, 0.08]$

Young women

- $\langle\{\text{education}\}\rangle$, $[1.84, 0.26]$
- $\langle\{\text{education}\}, \{\text{work}\}\rangle$, $[4.01, 0.1]$
Рис.: TPR-FPR for classification by generation via contiguous prefix-based patterns
### Old men

- \((\{\{\text{work}\}\}, \{\text{marriage, separation}\}, \{\text{education}\}), [13.52, 0.025])\)
- \((\{\{\text{work}\}\}, \{\text{marriage}\}, \{\text{separation}\}), [22.87, 0.042])\)
- \((\{\{\text{work}\}\}, \{\text{marriage}\}, \{\text{separation}\}, \{\text{education}\}), [\infty, 0.0208])\)

### Young men

- \((\{\{\text{education}\}\}, \{\text{work}\}, \{\text{separation}\}, \{\text{marriage}\}, \{\text{children}\}), [10.58, 0.020])\)
- \((\{\{\text{education}\}\}, \{\text{work}\}, \{\text{separation, partner}\}, \{\text{marriage}\}), [8.65, 0.016])\)
- \((\{\{\text{education}\}\}, \{\text{marriage, separation}\}), [7.69, 0.015])\)
We have studied several pattern mining techniques for demographic sequences including pattern-based classification in particular.

We have fitted existing approaches for sequence mining of a special type (contiguous and prefix-based ones).

The results for different demographic groups (classes) have been obtained and interpreted.

In particular, a classifier based on emerging sequences and pattern structures has been proposed.
Thank you!

Questions?