Temporal Skeletonization on Sequential Data
Patterns, Categorization, and Visualization

Chuanren Liu
Rutgers University
chuanren.liu@rutgers.edu

Chuanren Liu$^1$, Kai Zhang$^2$, Hui Xiong$^1$, Geoff Jiang$^2$, Qiang Yang$^3$

$^1$ Rutgers University
$^2$ NEC Laboratories America, Inc.
$^3$ Hong Kong University of Science and Technology
Diversified applications in dynamic business environments.

Challenge: determine the right granularity for sequential pattern mining with curse of cardinality.

Cardinality of sequential data: number of symbols.
Business-to-Business Purchase Pattern Analysis

Challenge: symbolic items with unknown stages.
Motivation: Curse of Cardinality

\[ \Pr(p) = c^{-\ell} \]

\( p \): pattern  \( c \): cardinality  \( \ell \): pattern length

Complexity

\[ m,h,j,f,d,a,i,k,b \]
\[ j,l,m,a,n,f,b,o,g \]
\[ e,h,l,c,f,n,i,b,o \]
\[ h,l,e,c,a,f,k,o,i \]

Granularity

\[ m,k,j,f,d,a,i,h,b \]
\[ j,l,m,a,n,f,b,o,g \]
\[ e,h,l,c,f,n,i,b,o \]
\[ h,l,e,c,m,f,k,o,i \]

Rareness

\[ m,k,j,f,d,a,i,h,b \]
\[ j,l,m,a,n,f,b,o,g \]
\[ e,h,l,c,f,n,i,b,o \]
\[ h,l,e,c,m,f,k,o,i \]

Noise
Existing Approaches

Figure: Taxonomy of symbols.

Table: Features of symbols.

Disadvantages:

- It is difficult to obtain knowledge of items.
- It is difficult to define distances and cluster the items.
- Performed irrespective of the temporal content.
- Cannot identify relevant temporal structure.
Our Approach: Temporal Skeletonization

Proactively reduce cardinality by temporal clusters:
- Summarize the temporal correlation in graph.
- Embed the graph in a low dimensional space.
- Identify temporal clusters.
- Re-encode sequences with temporal clusters.

(a) Random sequences  
(b) Seqs with temporal clusters  
(c) Customer event sequences

Figure: The embedding of items in different types of sequence data.
Graph Model for Temporal Skeletonization

Notations:

- Symbols $S = \{e_1, e_2, \cdots, e_{|S|}\}$.
- Sequence $S_n = (s^n_1, s^n_2, \cdots, s^n_{T_n})$.
- Sequences $\{S_n | n = 1, 2, \cdots, N\}$.

Objective: a coding scheme $y = f(e) \in \{1, 2, \cdots, K\}$ by

$$
\min \frac{1}{N} \sum_{n=1}^{N} \sum_{1 \leq p, q \leq T_n, |p - q| \leq r} (f(s^n_p) - f(s^n_q))^2
$$

Relax the integer constraints to:

$$
y_i = f(e_i) \in \mathbb{R}.
$$
Graph Model for Temporal Skeletonization II

Define the temporal graph $W$ as

$$W_{ij} = \frac{1}{N} \sum_{n=1}^{N} \sum_{1 \leq p, q \leq T_n} \sum_{|p-q| \leq r} [s^n_p = e_i \land s^n_q = e_j]$$

Then our objective is to minimize

$$\sum_{i,j} W_{ij} (y_i - y_j)^2.$$

This is an standard graph-based optimization problem (with constraints avoiding trivial solutions).
Embedding and Visualization

- Compute eigenvectors of graph Laplacian as embedding.
- Cluster the items in the embedding space.
- Transform raw sequences to sequences of temporal clusters.

**Figure:** The embedding of symbols in different types of sequence data.

(a) Random sequences  
(b) Seqs with temporal clusters  
(c) Customer event sequences
Post-Temporal-Smoothing

- **Objective**: Further reduce temporal variations in individual sequences.
- **Solution**: Gaussian mixture models with fused lasso.
- **Notations**: Given soft stages $Y \in \mathbb{R}^{T \times K}$, compute a smoother version $X^n$:
- **Optimization**: \[
\max \sum_{t=1}^{T} \sum_{k=1}^{K} X_{tk} Y_{tk}, \text{ subject to }
\frac{1}{T-1} \sum_{t=1}^{T-1} \|X_t - X_{t+1}\|_1 \leq \lambda, \\
\sum_{k=1}^{K} X_{tk} = 1, X_{tk} \geq 0
\]
Applications

1. **Sequence visualization**
   - Visualize symbol with coordinates.
   - Visualize sequence as a trajectory.

2. **Sequential pattern mining**
   - Temporal cluster as new granularity.
   - Curse of cardinality can be relieved.

3. **Sequence clustering**
   - Noises removed.
   - More meaningful features.
5000 sequences with:

- **Stages**: \(A, B, C, D, E\)
- **Patterns**: \((A \rightarrow B \rightarrow C \rightarrow D), (B \rightarrow E \rightarrow C)\)
- **Symbols**: 25 for each stage.

Simulation process:
- Determine the stage to sample from based on the pattern.
- Sample \(d\) symbols from the stage.

\[
d \sim (1 - p)p^{d-1}, \quad p = \frac{14}{15}, \quad \mathbb{E}[d] = 15.
\]
Using the temporal clusters (stages) identified via our approach, the mining process succeeds quickly in less than 1 second, and can recover exactly the two ground truth stage-wise patterns (when support is less than or equal to 0.5).

**Figure:** FSM algorithms on the simulated data.
Baselines II

HMM iterations:

1. For all sequences in cluster $C_k$, estimate a transition matrix $\phi_k$ and a emission matrix $\theta_k$;

2. Reallocate each sequence $S_n$ to the cluster $C_k$ on whose transition and emission matrices it has the highest probability of being produced, i.e., $k = \arg \max_k \Pr(S_n | \phi_k, \theta_k)$.

<table>
<thead>
<tr>
<th>Task</th>
<th>Sequence clustering</th>
<th>Stage recovery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>HMM</td>
<td>Ours</td>
</tr>
<tr>
<td>Precision</td>
<td>0.997</td>
<td>1</td>
</tr>
<tr>
<td>Recall</td>
<td>0.997</td>
<td>1</td>
</tr>
</tbody>
</table>

Table: Error of HMM and our method on two tasks.
Our Results

- We can perfectly recover the stages.
- We can perfectly cluster the sequences.
- We do not require prior knowledge on the data.

(a) Five symbol clusters

(b) Two sequence clusters
Data Description

Huge amount of customer event data:

- 88040 customers
- 5028 event symbols
- 248725 event records
Embedding Results and Buying Stages

<table>
<thead>
<tr>
<th>C</th>
<th>Top keywords</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>Official Website</td>
<td>12</td>
</tr>
<tr>
<td>C2</td>
<td>Corporate Event, Direct Mail</td>
<td>20</td>
</tr>
<tr>
<td>C3</td>
<td>Trial Product Download</td>
<td>45</td>
</tr>
<tr>
<td>C4</td>
<td>Conference</td>
<td>27</td>
</tr>
<tr>
<td>C5</td>
<td>Unsubscribe</td>
<td>38</td>
</tr>
<tr>
<td>C6</td>
<td>Webinar</td>
<td>101</td>
</tr>
<tr>
<td>C7</td>
<td>Trial Product Download</td>
<td>70</td>
</tr>
<tr>
<td>C8</td>
<td>Tradeshow</td>
<td>37</td>
</tr>
<tr>
<td>C9</td>
<td>Corporate Event, Direct Mail</td>
<td>65</td>
</tr>
<tr>
<td>C10</td>
<td>Web Marketing Ads</td>
<td>13</td>
</tr>
<tr>
<td>C11</td>
<td>Webinar</td>
<td>21</td>
</tr>
<tr>
<td>C12</td>
<td>Webinar</td>
<td>42</td>
</tr>
<tr>
<td>C13</td>
<td>Search Engine</td>
<td>12</td>
</tr>
</tbody>
</table>
Sequential Patterns in B2B customer event data

Figure: Sequential patterns in B2B customer event data.
Critical Buying Paths

<table>
<thead>
<tr>
<th>Class</th>
<th>P</th>
<th>Path/Keyword</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Successful</td>
<td>P1</td>
<td>$C_{10} \rightarrow C_1 \rightarrow C_7 \rightarrow C_{12} \rightarrow C_8$ Ads$\rightarrow$Website$\rightarrow$Download$\rightarrow$Webinar$\rightarrow$Tradeshow</td>
<td>933</td>
</tr>
<tr>
<td></td>
<td>P2</td>
<td>$C_{13} \rightarrow C_3 \rightarrow C_{11} \rightarrow C_{12} \rightarrow C_8$ Search$\rightarrow$Download$\rightarrow$Webinar$\rightarrow$Webinar$\rightarrow$Tradeshow</td>
<td>1110</td>
</tr>
<tr>
<td></td>
<td>P3</td>
<td>$C_6 \rightarrow C_{11} \rightarrow C_{12} \rightarrow C_8$ Webinar$\rightarrow$Webinar$\rightarrow$Webinar$\rightarrow$Tradeshow</td>
<td>702</td>
</tr>
<tr>
<td>Unsuccessful</td>
<td>P4</td>
<td>$C_2 \rightarrow C_9 \rightarrow C_5$ Mail$\rightarrow$Corporate Event$\rightarrow$Unsubscribe</td>
<td>423</td>
</tr>
<tr>
<td></td>
<td>P5</td>
<td>$C_{11} \rightarrow C_4 \rightarrow C_5$ Webinar$\rightarrow$Conference$\rightarrow$Unsubscribe</td>
<td>333</td>
</tr>
</tbody>
</table>
Comparison with baselines

<table>
<thead>
<tr>
<th>State</th>
<th>Top keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_1$</td>
<td>Seminar, Official Website, Trial Download</td>
</tr>
<tr>
<td>$H_2$</td>
<td>Seminar, Corporate Event, Tradeshow</td>
</tr>
<tr>
<td>$H_3$</td>
<td>Trial Download, Seminar, Corporate Event</td>
</tr>
<tr>
<td>$H_4$</td>
<td>Seminar, Conference, Corporate Event</td>
</tr>
<tr>
<td>$H_5$</td>
<td>Seminar, Unsubscribe, Corporate Event</td>
</tr>
</tbody>
</table>
Temporal skeletonization

- A new approach to address curse of cardinality:
  - Translate rich temporal content into topoly space.
  - Explore, quantify, and visualize temporal structures.

- Applied on B2B customer event data:
  - Discover critical purchasing patterns.
  - Identify dynamic buying stages of customers.
  - Improve the marketing practice.
Thanks!