Trajectory Data Mining in the Age of Big Data and AI

Ontario Database Day (OnDBD 2023)

Manos Papagelis

Wed, Dec 13, 2023
McMaster University
Background & Motivation
Trajectory/Mobility Data

Trajectory: A Sequence of (Spatiotemporal) Points

Vast Amounts of Trajectory/Mobility Data
Trajectory Data Mining

trajectory similarity

trajectory clustering

trajectory anomaly detection
trajectory network mining
trajectory classification

...
Trajectory Data Mining in the Age of Big Data and AI

A symbiotic relationship that presents a new strategy for addressing complex problems in trajectory data mining.

Image source: This image was created with the assistance of DALL·E 3
Plethora of Applications

- ridesharing
- trip/POI (point-of-interest) recommendation
- traffic analysis
- route planning and optimization
Data Mining Lab @ YorkU’s Journey on Trajectory Data Mining

- Trajectory simplification [ACM SIGSPATIAL ‘23]
- Trajectory similarity [Submitted]
- Trajectory dataset and resources [ACM SIGSPATIAL ‘23]
- Trajectory prediction [Submitted]
- Trajectory classification [IEEE MDM ‘23]

- Trajectory network analysis [Big Data Research, IEEE MDM ’20, Geoinformatica, IEEE BigData ‘18, 2 x IEEE MDM ’18]
- Mobility + epidemics [ACM SIGSPATIAL/SpatialEpi ’24, ACM SIGSPATIAL/SpatialEpi ’23, IEEE MDM ’22]
- Transportation optimization [ACM SIGSPATIAL ’22, ACM SIGSPATIAL ’22]
Trajectory Simplification

The Trajectory Pathlet Dictionary Construction Problem
Trajectories on the Road Network

- **Trajectory**
  - Denoted by $\tau$
  - Represented as:
    $$\tau = \langle (x_1, y_1, t_1), \ldots, (x_{|\tau|}, y_{|\tau|}, t_{|\tau|}) \rangle$$

- **Road Network**
  Modelled as a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$
  - $\mathcal{V}$: **Nodes** (set of road intersections)
  - $\mathcal{E}$: **Edges** (set of road segments)
Road Segment-based Representation

- Each trajectory \( \tau \) can be expressed as a set of road segments \( R_s \subseteq R \)
- This special representation is denoted by \( \mathcal{R}(\tau) \)
Trajectory Pathlet Dictionary (PD) Construction

• **Pathlet Dictionary**: A small set of basic building blocks that can represent a wide range of trajectories

• Many names in the literature


  • Pathlet
  • Subtrajectory
  • Trajectory Segments
  • Fragments
  • …
Pathlet-based Representation of a Trajectory

Denoted by $\Phi(\tau) = \{\rho^{(1)}, \rho^{(2)}, \ldots, \rho^{(k)}\}$

\[
\Phi(\tau) = \{\rho_1, \rho_5, \rho_6, \rho_3\}
\]
PathletRL - Overview

- Extracting candidate pathlets
- Deep Reinforcement Learning framework
Extracting Candidate Pathlets - Example

Choose pathlet $\rho$ uniformly at random
Identify all neighbors of $\rho$
Compute utility of $\rho_{merge}$
Obtain $\rho^*$ with the highest utility
Merge $\rho$ and $\rho_4$
New current pathlet $\rho$

<table>
<thead>
<tr>
<th>$\rho_{merge}$</th>
<th>Utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>MERGED($\rho, \rho_1$)</td>
<td>+0.7</td>
</tr>
<tr>
<td>MERGED($\rho, \rho_2$)</td>
<td>+1.8</td>
</tr>
<tr>
<td>MERGED($\rho, \rho_3$)</td>
<td>-1.6</td>
</tr>
<tr>
<td><strong>MERGED($\rho, \rho_4$)</strong></td>
<td><strong>+5.5</strong></td>
</tr>
<tr>
<td>MERGED($\rho, \rho_5$)</td>
<td>-3.2</td>
</tr>
<tr>
<td>MERGED($\rho, \rho_6$)</td>
<td>+2.9</td>
</tr>
</tbody>
</table>
Approach & Contributions

Edge-disjoint pathlets

Deep Reinforcement Learning (DQN)

Partial trajectory reconstruction ~85%
Trajectory Similarity

The Top-k Trajectory Similarity Search Problem
Trajectory Similarity

- How similar two trajectories are
- Several ways to define
Spatiotemporal Similarity – Example

(a) Taxi Trajectories

(b) Embedding Space

proximate each other
Problem Statement

- **Top-\(k\) Trajectory Similarity Search Task**
  - **Given**: Trajectory set \(\mathcal{T}\)
    - Query trajectory \(\tau_q\)
    - Positive integer \(k \geq 1\)
  - **Find** the (ranked) list of top \(k\) trajectories in \(\mathcal{T}\):
  - **Criterion**: Similarity with \(\tau_q\)

1: \(\tau^{(1)}\)
2: \(\tau^{(2)}\)
3: \(\tau^{(3)}\)
...
k: \(\tau^{(k)}\)
Approach: Reducing Trajectory Similarity to Set Similarity Problem

- Treat each trajectory as a set; its elements are the road segments it has traversed (road-based representation $\mathcal{R}(\tau)$)
- Similar (Dissimilar) trajectories map to similar (dissimilar) sets

Trajectories $\tau_1$ and $\tau_2$ are similar!
ST2Box Overview

- Spatiotemporal Trajectories to Box Embeddings for Similarity Learning
ST2Box Overview

- Spatiotemporal Trajectories to Box Embeddings for Similarity Learning

- Seg2Vec – spatiotemporal vector representations of road segments
ST2Box Overview

- Spatiotemporal Trajectories to Box Embeddings for Similarity Learning

- Seg2Box – box representations of sets of road segments
ST2Box Overview

• Spatiotemporal Trajectories to Box Embeddings for Similarity Learning

• Overlapping boxes → Similar sets → Similar trajectories

ST2Box Properties

Accurate, Versatile, Generalizable, Robust, Fast, Scalable

Up to ~30% Performance Gain
Higher-order Mobility Flow Data
Map Tessellation

lower resolution

higher resolution
Trajectories: Sequences of Hexagons

Trajectory: \( h_1, h_2, h_3 \ldots h_{20}, h_{21}, h_{22} \)
Treat Trajectories as Language Statements

Hexagons represent ‘tokens’ & trajectories represent ‘sentences’

Trajectory:  

Sentence:  I like to learn English

Advantages:

- Reduced data sparsity
- More compatible with well-known ML models (e.g., sequence models, LLMs)
Point2Hex: Overview of the Pipeline

- GPS Traces or POI Check-Ins (input)
- Linestring of Trajectories (Map-matching)
- Map Tessellation with Trajectories (Hexagon-shaped cells)
- Intersection of Linestrings and Polygons (Computational Geometry)
- Higher-order Mobility Flow (Output)
### Higher-order Mobility Flow: Datasets and Data Generator

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Trajectories</th>
<th>Time Period</th>
<th>Resolutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>HO-T-Drive</td>
<td>65,117</td>
<td>02/02/08 - 02/08/08</td>
<td>{6,...10}</td>
</tr>
<tr>
<td>HO-Porto</td>
<td>1,668,859</td>
<td>07/01/13 - 06/30/14</td>
<td>{6,...10}</td>
</tr>
<tr>
<td>HO-Rome</td>
<td>5,873</td>
<td>02/01/14 - 03/02/14</td>
<td>{6,...10}</td>
</tr>
<tr>
<td>HO-GeoLife</td>
<td>2,100</td>
<td>04/01/07 - 10/31/11</td>
<td>{6,...10}</td>
</tr>
<tr>
<td>HO-FourSquare-NYC</td>
<td>49,983</td>
<td>04/12/12 - 02/16/13</td>
<td>{6,...10}</td>
</tr>
<tr>
<td>HO-FourSquare-TKY</td>
<td>117,593</td>
<td>04/12/12 - 02/16/13</td>
<td>{6,...10}</td>
</tr>
<tr>
<td>HO-NYC-Taxi</td>
<td>2,062,554</td>
<td>01/01/16 - 06/30/16</td>
<td>{6,...10}</td>
</tr>
</tbody>
</table>

Datasets @ [Zenodo](#)

Data Generator @ [GitHub](#)
Trajectory Prediction

Predict the Next-k Trajectory Steps Problem
Problem of Interest: Trajectory Prediction

History trajectory

Predict future trajectory
Trajectory Prediction (Revisited)

Let
- an observation area
- a set of objects and their history trajectories
- an observation period

Input: Given
- a moving object $n$
- a partial trajectory $= <p_1, p_2, ..., p_t>$
- a prediction horizon $k > 0$

Output: We want to predict the next $k$ hexagons of the input partial trajectory
Approach & Contributions

Trajectory Prediction (Revisited)

**TrajLearn**: Trajectory Deep Generative Model

Beam search
Trajectory Classification
The Trajectory-User Linking Problem
can trajectories help to identify a person?
Trajectory-user Linking (TUL)

trajectory-user linking aims at linking anonymous trajectories to users who generate them
Problem Definition

Trajectory-user linking aims at linking anonymous trajectories to users.

Given:
\[ \mathcal{U} = \{u_1, u_2, u_3, \ldots, u_c\} \] – users
\[ \mathcal{T} = \{Tr_1, Tr_2, \ldots, Tr_n\} \] – unlinked trajectories

TUL is defined as a multiclass classification problem

\[
\min_{f \in \mathcal{F}} \mathbb{E}[\mathcal{L}(f(Tr_i), u_i)] \text{ over } \mathcal{F}
\]

where \( \mathcal{F} \) is the set of all classifiers in the hypothesis space
\( \mathcal{L}(\cdot) \) is the loss between the predicted label \( f(Tr_i) \in \mathcal{U} \) and the true label \( u_i \in \mathcal{U} \)
Higher-order mobility flow data generation

**TULHOR**: A spatiotemporal model that deals with sparsity and low data quality of the TUL problem

**TULHOR** outperforms baselines by up to 8%
Questions
Credits

Gian Alix

Mahmoud Alsaeed

Ali Faraji

Jing Li

Nina Yanin

Amirhossein Nadiri


Thank you!