Spatial Trajectory Analytics: Past, Present and Future

Xiaofang Zhou

Trajectory Data
...data about moving objects

What is Trajectory Data
- Any data that record the locations of a moving object over time in a geographical space

Simple form:
\(<ID, (p_1,t_1), (p_2,t_2), ... (p_n,t_n)\>
ordered by time: \(t_1 < t_2 < ... < t_n\)

General form:
\(<oID, tID, trajProperties, (p_1,t_1,a_1), (p_2,t_2,a_2), ... (p_n,t_n,a_n)\>

Massive Amount of GPS Data

Other Types of Trajectory Data

SENORS

More about Trajectory Data
- A trajectory is obtained from sampling the movement of an object
- Some sampling strategies are used → not only data, but also models to generate data
- Objects movement with constraints (e.g., by map) → not only data, but also environment data
- There are many other factors which cannot be controlled → data quality issues
- Data can be both redundant as well as sparse → compression, alignment and prediction
- It is non-trivial even to restore the original trace from a trajectory → harder to compare
Are Trajectory Data Useful?

- Route planning
- POI recommendation
- LBS and mobile advertisement
- Resource tracking and scheduling
- Fleet management
- Road safety
- Emergency responses
- Environment monitoring...
- Urban planning and smart cities

Trajectory analytics now becomes a new frontier for business intelligence, especially in real-time and in combination with other types of data.

Processing Trajectory Data

...monitoring, managing and processing

Trajectory Processing Framework

The Past

...driven by curiosity

Moving Objects/Trajectory Work

- Initially on foundations
  - Data representation, query languages and basic operations, indexing methods etc.
- Curiosity-driven
  - Imagine a special “novel” type of query, find a “novel” indexing method and then use “standard” methods to improve efficiency
- Not directly useful
  - Strong assumptions (not useful in practice)
  - Highly specialized indexes (cannot be implemented)
- Also, data mining, social networks, recommender...

An Introduction Book

- Computing with Spatial Trajectories
  - Yu Zheng and Xiaofang Zhou, 2011
- Part I Foundations
  - Trajectory Preprocessing (W.C. Lee, J. Krumm)
  - Trajectory Indexing and Retrieval (Y. Zhou et al)
- Part II Advanced Topics
  - Uncertainty in Spatial Trajectories (G. Trajcevski)
  - Privacy of Spatial Trajectories (C. Y. Chow, M. Mokbel)
  - Trajectory Pattern Mining (Y. Zhu, Y. Zheng, G. Yang)
  - Activity Recognition from Trajectory Data (Y. Zhu, Y. Zheng, G. Yang)
  - Trajectory Analysis for Driving (J. Krumm)
  - Location-Based Social Networks: Users (Y. Zheng)
  - Location-Based Social Networks: Locations (Y. Zheng and X. Xie)
The Present
...continues the past and also driven by new trends

Trajectory Data Research: Trends

Trajectory Research: Topics

Trajectory Research: What’s New

A Systematic View

Some of Our Work

Prediction of movement [ICDE08] and paths [VLDBJ10], trajectory simplification with error bound [VLDB08], path nearest neighbor query [SIGMOD09], searching trajectory by locations [SIGMOD10], most popular routes [ICDE11], probabilistic range query [EDBT11, ICDE12], materialized shortest paths [TOIS12], spatial keyword search for trajectories [ICDE13,15,16], clue-based queries [VLDB17], minimum on-road time routing [VLDB17, VLDBJ18], trajectory calibration [SIGMOD13, VLDBJ15], route and location recommendation [ICDE14, SIGKDD15, ICDE16, TOIS16, TIST18], trajectory exploration and summarization [ICDE15], in-memory trajectory databases [CIKM14, SIGMOD15], privacy-preserving trajectory search [ICDE15], data sparsity [MDM18], ML for speed prediction [IJCAI18]
Minimum On-Road Time Route Planning
...yet another “new” type of queries

Time Dependent Path
- A Path $p =\langle v_1, \ldots, v_k \rangle$ from $v_1$ to $v_k$
  - $a(v_i)$: arrival time
  - $\beta(v_i)$: departure time
  - $\forall v_i \in V$: $a(v_i) - \beta(v_i) \geq 0$
- Total Travel Time
  - $\text{Travel} = a(v_k) - \beta(v_1)$
- Fastest Path
  - $\min(a(v_k) - \beta(v_i))$
  - Because of FIFO, only when $\beta(v_i) - a(v_i)$>0 can lead to optimal solution

Fastest Path
- Single Starting-Time Fastest Path
  - $Q((v_0, v_k), t_0)$
    - Starting time $t_0$ on $v_0$ is fixed
    - Can use Dijkstra directly [1]
  - Time Complexity: $O(|V|\log |V| + |E|)$

Variations
- Earliest Arrival Path
- Min$(\delta(v_0))$
- Same as single starting-time fastest path
- Latest Departure Path
- Max$(\delta(v_k))$
- Reverse

Interval Starting-Time Fastest Path
- $Q([t_1, t_2], t_0)$
  - $t_1$: starting time interval
  - $t_2$: latest arrival time
  - Min$(\delta(v_0))$
- Find the optimal starting time within $[t_1, t_2]$
  - Solution in [1]
    - Only waiting on $v_1$ is considered.
- Problem Complexity
  - $O(T(|V|\log |V| + |E|))$ [2]
  - If $a$ is linear piecewise functions, $T$ is the turning points in the final result

But...
- Is the Fastest Path always best?
  - For Logistics company
    - Major operation cost -- Fuel consumption -- Time spent on road
  - For tourists
    - Reduce their time spent on road so that they can spend more time on the tourist attractions.

- Spending less time on road rather than arriving final destination earlier

Time Dependent Graph
- Time-Dependent Graph is defined as $G_T(V, E, W)$
  - $V = \{v_0, \ldots, v_k\}$ vertex set
  - $E \subseteq V \times V$ directed edge set
  - $W$ a set of cost functions
    - $w(v_i, v_j, t) \in W$
    - Tells how much time it costs to travel from $v_i$ to $v_j$ at time $t$
  - Also known as Speed Profile
- Can Dijkstra algorithm still work?
  - FIFO Time-Dependent Graph
    - Given a time dependent graph $G_T(V, E, W)$, it is a FIFO Graph if
      $w(v_i, v_j, t) \leq w(v_i, v_j, t')$, when $t > t'$.

---

Naïve Approach

1. Find the fastest path from $v_i$ to $v_d$ $< v_i, v_{p+1}, v_{p+2}, ... , v_d >$ with the optimal starting time $t_o$ on $v_i$
2. Find the fastest path from $v_{p+1}$ to $v_d$ $< v_{p+1}, v_{p+2}, ... , v_d >$ with optimal starting time $t_{p+1}$ on $v_{p+1}$
3. Run repeatedly at every node

Any problems?

On-Road Time

- A Path $p =< v_1, ..., v_k >$ from $v_1$ to $v_k$
  - $a(v_i)$: arrival time
  - $t(v_i)$: departure time
  - $v_i \in V - V'$: $t(v_i) = a(v_i)$
  - $v_i \in V'$: $t(v_i) \geq a(v_i)$

On-road Travel Time

- $\text{Wait} = \sum_{i=1}^{l} (w(v_i, v_j) \cdot t(v_i, j))$

Time Dependent Path Problems

<table>
<thead>
<tr>
<th>Graph Type</th>
<th>Path Problem</th>
<th>Objective</th>
<th>Waiting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static Graph</td>
<td>Shortest Path</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>Time Dependent Graph</td>
<td>Single static time fastest path</td>
<td>(w(V_t) = \sum_{i=1}^{l} w(v_i, v_j) \cdot t(v_i, j))</td>
<td>Total Static Weight</td>
</tr>
<tr>
<td></td>
<td>General</td>
<td>$w(V_t) = \max_{i=1}^{l} (w(v_i, v_j) \cdot t(v_i, j))$</td>
<td>Total Weight</td>
</tr>
<tr>
<td></td>
<td>Earliest</td>
<td>(w(V_t) = \min_{i=1}^{l} (w(v_i, v_j) \cdot t(v_i, j)))</td>
<td>Total Weight</td>
</tr>
<tr>
<td></td>
<td>Latest Dependent</td>
<td>(w(V_t) = \max_{i=1}^{l} (w(v_i, v_j) \cdot t(v_i, j)))</td>
<td>Total Weight</td>
</tr>
<tr>
<td></td>
<td>Interval start time fastest path</td>
<td>(w(V_t) = \min_{i=1}^{l} (w(v_i, v_j) \cdot t(v_i, j)))</td>
<td>Total Weight</td>
</tr>
<tr>
<td></td>
<td>Minimum Total Time</td>
<td>(w(V_t) = \min_{i=1}^{l} (w(v_i, v_j) \cdot t(v_i, j)))</td>
<td>Total Weight</td>
</tr>
<tr>
<td></td>
<td>Solution Vertex</td>
<td>A set of parking vertices</td>
<td>False</td>
</tr>
</tbody>
</table>

Algorithm Overview

- Problem: Find a path with $\min \left( \sum_{i=1}^{l} w(v_i, v_j, t(v_j)) \right)$
- MORT Algorithm
  1. Active Time Interval
  2. Path Expansion
    - $C_t(t)$: The minimum on-road travel time from $v_i$ to $v_j$ that arrives $v_j$ on $t$.
    - Update $C_t(t)$ for all $v_i$ until the optimal is reached.
    - Non-increasing property of parking vertices
    - Waiting will increase the total travel time.
    - Waiting will not increase the on-road travel time.
  3. Incremental Algorithm: $\delta \left( C_t(t) \big| \gamma \left( t \right) + \delta \left( t \right) \right)$
  4. Approximation
  5. Route Retrieval

L. Li, S. Xue, X. Du and Y. Zhao, “Minimal On-Road Time Route Scheduling on Time-Dependent Graph,” VLDB 2017.
Road Speed Profile Generation

...another example of large-scale space problem

Speed Profile Generation

1. Map Matching
   - Attach the GPS points on the roads
   - Use the length and temporal information to get speed of each road at different time
2. Speed Data Collection
   - 5-Minute Histogram
3. Missing Value Estimation
   - Cosine Similarity, Matrix Factorization based Collaborative Filtering, HMM...
4. Compression
   - PLA (Piecewise Linear Approximation)


Data Sparsity

...no matter how much data you have, you don’t have enough

Trajectory Clustering and Labeling

Applications
- Moving Behaviors Analysis
- Personalized Routing

Clustering
- Data Sparsity: Origin-Destination Trajectory
- Clusters Identification: Efficient Algorithms

Labeling
- Features: Fastest, Shortest, Most popular, Time-dependent

Trajectory Augmentation

- Data augmentation approach
  - Factorization-based [1]: extra data sources (geo-spatial, temporal, and historical correlation)
  - Concatenation-based [2]: sub-trajectories

No Robustness Validation Guarantee

Deep-Learning for Speed Predication

...many things we do is about using spatial data for prediction
Deep Learning + ITS = ?

The Problem

- Given:
  - A road map (as a directed graph)
  - A sequence of speed vectors, each vector is the speed at each road segment during a time interval \( X_t = [x_{t_1}, x_{t_2}, \ldots, x_{t_n}] \)

- Problem: Given the historical observations \( \{X_t\} = 1, \ldots, t \), this paper aims to predict \( Y_t = \{X_{t+1}, \ldots, t+z\} \), where \( z \) is the number of time intervals to be predicted.

Challenges

- Spatiotemporal characteristics
  - Speed on a road tends to slow down if congestions occur in its surrounding area
  - Speed in the downtown area is relatively low during rush hours
- Restricted by road network
  - Traffic in a certain road just affects its adjacent roads
  - Traditional CNN cannot capture topology of road network
- Other factors
  - Context Information: e.g., weather, holiday
  - Periodic Law: e.g., daily period, weekly trend

Generating Speed Vectors

LC-RNN Model

- Look-up Convolution (LC): learn the latent features of surrounding area
- LSTM components: learn the time-series pattern that is aware of surrounding area dynamics

Results

- Beijing Dataset
  - A big road network about 10 thousand main roads
  - Trajectory data comes from 1 Mar to 31 Jul in 2016
  - The first 4 months as train and the last 1 month as test

- Shanghai Dataset
  - A small road network about 1.5 thousand main roads
  - Trajectory data comes from 1 Mar to 31 Apr in 2015
  - The first 45 days as train and the last 15 days as test

The Future

...driven by real data and real problems

New Context

- Much more data
- Large-scale space problem
- Dynamic environment
- Much more queries
- Personalization
- Privacy concerns
- Machine learning
- ...

More Data

...more things can be done, but need a scalable and general purpose tool?

Trajectory Data in a Company

- A car navigation service provider
- Total trajectory data: 32 TB in size, 10.9 billion matched trajectories

<table>
<thead>
<tr>
<th>Company</th>
<th>Current (TB)</th>
<th>Daily (trajectories)</th>
</tr>
</thead>
<tbody>
<tr>
<td>X (in-car navigation provider)</td>
<td>17.8</td>
<td>10M</td>
</tr>
<tr>
<td>Y (map app provider)</td>
<td>14.5</td>
<td>5M</td>
</tr>
<tr>
<td>Z (social network)</td>
<td>0.68</td>
<td>1M</td>
</tr>
</tbody>
</table>

- Every day ~40M new trajectories, ~4 billion points
- Sampling rates: 50% ~2s, 99% < 10s

NavInfo DataHIVE (minedata.cn)

<table>
<thead>
<tr>
<th>Vehicle</th>
<th>Infrastructure</th>
<th>Environment</th>
<th>People</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trajectories:</td>
<td>Standard maps</td>
<td>Weather</td>
<td>Voice and text</td>
</tr>
<tr>
<td>- taxis</td>
<td>High res maps</td>
<td>Events</td>
<td>User comments</td>
</tr>
<tr>
<td>- uber-like</td>
<td>Services POIs</td>
<td>Air quality</td>
<td>Search log</td>
</tr>
<tr>
<td>- monitored</td>
<td>Culture POIs</td>
<td>Water quality</td>
<td>Travel log</td>
</tr>
<tr>
<td>- commercial</td>
<td>Commercial POIs</td>
<td>Land &amp; water info</td>
<td>Operators' OD</td>
</tr>
<tr>
<td>- user-generated</td>
<td>Health POIs</td>
<td>DEM &amp; EEC</td>
<td>Workplace info</td>
</tr>
<tr>
<td>Sensor/OBD data</td>
<td>Travel POIs</td>
<td>Satellite image</td>
<td></td>
</tr>
<tr>
<td>Perception data</td>
<td>City models</td>
<td>Street views</td>
<td></td>
</tr>
<tr>
<td></td>
<td>City 3D Models</td>
<td>Roadside pictures</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Business districts</td>
<td>Laser point cloud</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Admin boundaries</td>
<td>Road condition</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Organization maps</td>
<td>Traffic condition</td>
<td></td>
</tr>
</tbody>
</table>

How Much?
Trajectory Data Management System
...a common platform and API

+ Why Common Platform?
  - Universal
    - GPS, telecom tokens, social apps...
  - Shared enterprise data
    - For monitoring, prediction, business insights...
  - Separation of conceptual, logical and physical design
    - Especially many computing platforms to consider today
  - Other benefits we took for granted
    - Optimization for data storage and query processing, scheduling, concurrency control...

+ The 5R Approach

+ A Spatiotemporal Pyramid

+ Trajectory Calibration
  - Popular trajectory distance measures
    - Euclidean distance
    - LCSS (longest common sequence)
    - DTW (Dynamic time warping)
    - EDR (Edit distance on Real sequences)
  - How distance measures work?
    - Sample points alignment
    - Aggregating differences of aligned pairs
  - Experiments
    - Ground Truth: 11,000 high-sampling-rate real trajectories
    - Derived Trajectory Datasets: re-sampling, shifting, jumping
    - Results?

+ SparkDB
  - A time-centric storage and processing system for trajectories
  - Designed for in-memory computers
  - A more ambitious system called Traminer is under development, following the proposed processing framework
  - Now supported by several companies

H. Su, K. Zhang, H. Wang and X. Zhou, Calibrating Trajectory Data for Similarity-based Analysis, SIGMOD 2013
Batch Fastest Path Queries

...can we do better if we can queries in batch?

**A Real Problem**
- At any time, 100K-1M OD pairs are given for route planning
- Options:
  - Processing them in parallel
  - Materializing all-pair shortest path information
  - Batch processing?
- Additional dimensions:
  - Ridesharing
  - Streaming requests (requests come continuously and cars are moving)

**Batch Shortest Path Queries**
- Case of 1:N
  - Dijkstra can be used straightaway
  - A* can be generalized
  - A good partition of N can improve efficiency
- Case of M:1
  - This can be done by reversing the above case
- Case of M:N
  - One-directional
  - Bidirectional
  - Partition-based

**Trajectory-based Entity Linking**

...everyone's movement pattern is unique

**People Have many Trajectory Datasets**
- Broadly available human mobility data:
  - Check-ins
  - Credit card transactions
  - Phone call records
  - Go-card records
  - Vehicle trajectories
  - Many social networks...

**Entity Resolution using Trajectory Data?**
- Understanding the extent to which spatiotemporal data are distinctive is crucial to:
  - Location privacy protection
  - Entity resolution
  - Data integration
  - Spatial content retrieval
  - Personalized location recommendation
  - Driver performance assessment
  - ...
Only 4 Points, Really?


"four spatiotemporal points are enough to uniquely identify 95% of the individuals"

Yes!

Signature: TFIDF-weighted point set
- Many possible reduction methods (simple frequency-based truncating, PCA, LSH)
- Several alternatives (may need different similarity measures)
- Cosine similarity
- Tested with 11K taxi data over 3 months in 3 major cities
- Accuracy: now extremely high
- Performance is bad: we are improving it

Top K | Precision
---|---
1 | 0.95987964
2 | 0.97241725
3 | 0.9779338
4 | 0.98194584
5 | 0.98445336
6 | 0.98546637
7 | 0.98690083
8 | 0.98746239
9 | 0.98845654
10 | 0.98898069

Some Other New Research Problems

...from discussions with companies doing trajectory analytics as their bread-and-butter business

A List of Problems

- ETA: O-side and D-side
- Map matching + map inferencing: an integrated approach
- Data fusion: among trajectory datasets and with others
- Similarity based search
- Traffic prediction: for prevention and intervention
- Transport mode detection: both large/small scale
- Personalized/constrained routing algorithms
- Privacy: can you really protect trajectory privacy?
- Smart city – a holistic traffic solution
- ...

Conclusions

- New problems
  - More data, more queries, more applications, more tools
  - From SDBMS, spatial data mining to spatial learning
- Some current research problems
  - Large-scale space problems
  - Dynamic road networks
  - Massive batch queries
  - Personalization and privacy issues
- We need a DBMS approach!
Thanks for listening

Q&A