Spatial Trajectory Analytics: Past, Present and Future

Xiaofang Zhou

The University Of Queensland AUSTRALIA

dke DATA & KNOWLEDGE ENGINEERING

Trajectory Data

…data about moving objects

+ What is Trajectory Data

3

- 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: $t_1 < t_2 < \ldots < t_n$
- 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

4

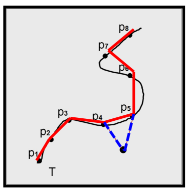
+ Other Types of Trajectory Data

5

+ More about Trajectory Data

6

- 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

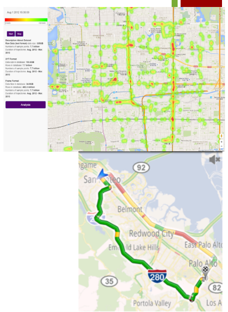


- It is non-trivial even to restore the original trace from a trajectory → harder to compare

+ Are Trajectory Data Useful?

7

- 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



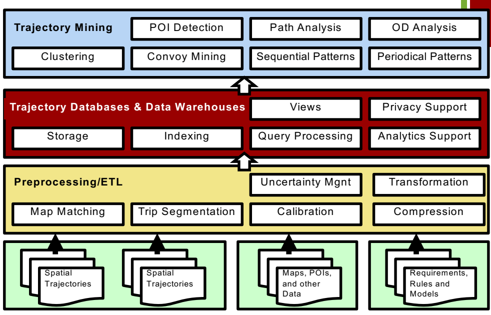
Trajectory analytics now becomes a new frontier for business intelligence, especially in real-time and in combination with other types of data

Processing Trajectory Data

…monitoring, managing and processing

+ Trajectory Processing Framework

9



The Past

…driven by curiosity

+ Moving Objects/Trajectory Work

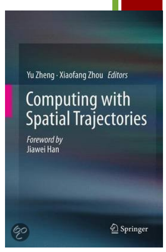
11

- 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

12

- ***Computing with Spatial Trajectories***
  - Yu Zheng and Xiaofang Zhou, 2011
- Part I Foundations
  - Trajectory Preprocessing *(W.-C. Lee, J. Krumm)*
  - Trajectory Indexing and Retrieval *(X. 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 *(H. Jeung, K. L. Yiu, C. Jensen)*
  - Activity Recognition from Trajectory Data *(Y. Zhu, V. Zheng, Q. Yang)*
  - Trajectory Analysis for Driving *(J. Krumm)*
  - Location-Based Social Networks: Users *(Y. Zheng)*
  - Location-Based Social Networks: Locations *(Y. Zheng and X. Xie)*

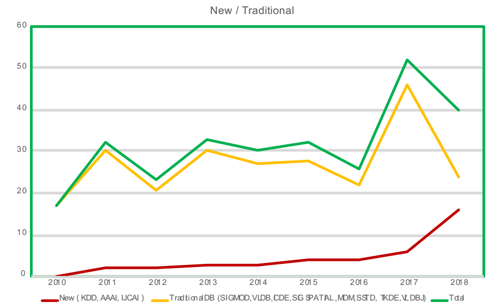


The Present

…continues the past and also driven by new trends

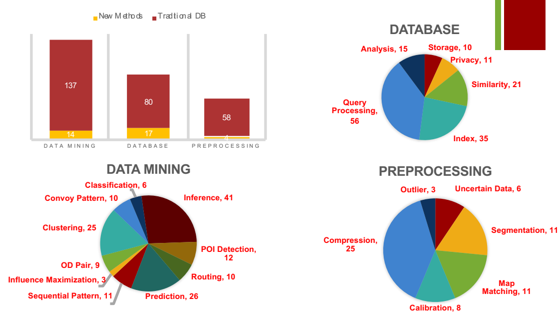
+ Trajectory Data Research: Trends

14



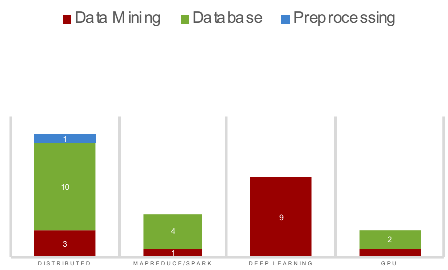
+ Trajectory Research: Topics

15



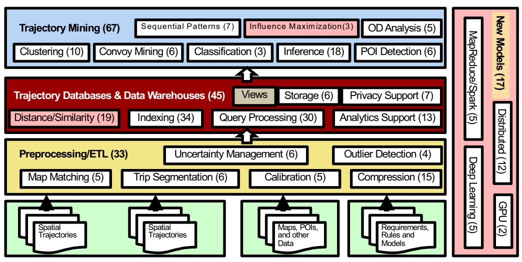
+ Trajectory Research: What's New

16



+ A Systematic View

17



+ Some of Our Work

18

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 [TODS12], 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 Graph

20

- Time-Dependent Graph is defined as $G_T\ (V, E, W)$
  - $V = \{v_i\}$ vertex set
  - $E \subseteq V \times V$ directed edge set
  - W a set of cost functions
    - $\forall (v_i, v_j) \in E,\ w(v_i, v_j, t) \in W$
    - 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 iff $\forall (v_i, v_j) \in E, w\ (v_i, v_j, t) \leq \Delta + w\ (v_i, v_j, t + \Delta)$, $where\ \Delta > 0$.

+ Time Dependent Path

21

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

+ Fastest Path

22

- Single Starting-Time Fastest Path
  - $Q(v_s, v_d, t_s)$
  - Starting time t_s on v_s is fixed
  - Can use Dijkstra directly [1]
  - Time Complexity: $O(|V|\log|V| + |E|)$
- Variations
  - Earliest Arrival Path
    - $Min(\alpha(v_d))$
    - Same as single starting-time fastest path
  - Latest Departure Path
    - $Max(\beta(v_s))$
    - Reverse

[1] Ding B, Yu J X, Qin L. Finding time-dependent shortest paths over large graphs. EDBT 2008

+ Fastest Path

23

- Interval Starting-Time Fastest Path
  - $Q(v_s, t_{s_1}, t_{s_2}, v_d, t_d)$
    - $[t_{s_1}, t_{s_2}]$: starting time interval
    - t_d latest arrival time
    - $Min(\alpha(v_d) - \beta(v_s))$
  - Find the optimal starting time within $[t_{s_1}, t_{s_2}]$
    - Solution in [1]
    - Only waiting on v_s is considered.
  - Problem Complexity
    - $\Omega(T(|V|\log|V| + |E|))$ [2]
    - If w is linear piecewise functions, T is the turning points in the final result

[1] B Ding, J X Yu, L Qin. Finding time-dependent shortest paths over large graphs. EDBT 2008
[2] L Foschini, J Hershberger, S Suri, On the complexity of time-dependent shortest paths, Algorithmica 2014

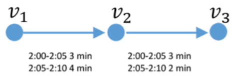
+ But…

24

- 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

+ Example

25



Fastest Path:

$V_1 \to V_2$ from 2:00 arrives at 2:03

$V_2 \to V_3$ from 2:03 arrives at 2:06

Total Time: 6 min On Road Time: 6 min

Minimal On-Road Time Path:

$V_1 \to V_2$ from 2:00 arrives at 2:03

Wait on V_2 until 2:05

$V_2 \to V_3$ from 2:05 arrives at 2:07

Total Time: 7 min On Road Time: **5 min**

+ Time Dependent Graph with Parking Nodes

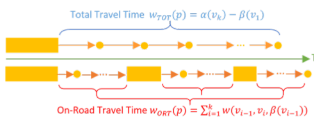
26

- Time-Dependent Graph is defined as $G_T\ (V, E, W)$
  - $V = \{v_i\}$ vertex set
  - $V' \subseteq V$, parking vertex set
  - $E \subseteq V \times V$ directed edge set
  - W a set of cost functions

+ On-Road Time

27

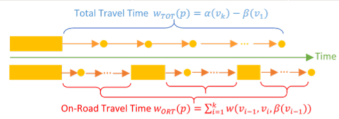
- A Path $p = < v_1, \ldots, v_k >$ from v_1 to v_k
  - $\alpha(v_i)$: *arrival time*
  - $\beta(v_i)$: *departure time*
  - $\forall v_i \in V - V'$: $\beta(v_i) = \alpha(v_i)$
  - $\forall v_i \in V'$: $\beta(v_i) \geq \alpha(v_i)$
- On-road Travel Time
  - $w_{ORT} = \sum_{i=1}^{k} (w(v_{i-1}, v_i, \beta(v_{i-1})))$



+ Time-Dependent Path Problems

28

Graph Type	Path Problem		Objective		Waiting
Static Graph	Shortest Path		$\sum_{i=s}^{d} Weight(p_{i,i+1})$	Total Static Weight	No
Time Dependent Graph	Single start time fastest path	General	$\sum_{i=s}^{d} Time(p_{i,i+1})$	Total Temporal Weight	No
		Earliest Arrival	$Min(Arrival(v_d))$		
		Latest Departure	$Max(Depart(v_s))$		
	Interval start time fastest path		$Min(Arrival(v_d) - Depart(v_s))$	Minimum Total Time	Source Vertex $V' = \{v_s\}$
	MORT		$Min(\sum_{i=1}^{k-1} w(v_{i-1}, v_i, \beta(v_{i-1})))$	Minimum On-Road Time	A set of parking vertices $V' \subseteq V$



+ Naïve Approach

29

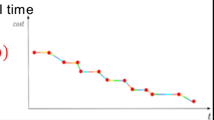
1. Find the fastest path from v_s to v_d $< v_s, v_{s+1}, v_{s+2}, \ldots, v_d >$ with the optimal starting time t_s on v_s
2. Find the fastest path from v_{s+1} to v_d $< v_{s+1}, v'_{s+2}, \ldots, v_d >$ with optimal starting time t_{s+1} on v_{s+1}
3. Run repeatedly at every node

Any problems?

+ Algorithm Overview

30

- Problem: Find a path with $Min(\sum_{i=1}^{k-1} w(v_{i-1}, v_i, \beta(v_{i-1})))$
- MORT Algorithm
  1. Active Time Interval
  2. Path Expansion
     - $C_i(t)$ The minimum on-road travel time from v_s to v_i that arrives v_i on t.
     - Update $C_i(t)$ 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
     1. Basic Algorithm: $O(T|V|\log|V| + T^2|E|)$
     2. Incremental Algorithm: $O(L(|V|\log|V| + |E|))$
     3. Approximation
  3. Route Retrieval



L. Li, W. Hua, X. Du and X. Zhou, "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

32

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)

L. Li, K. Zheng, X. Zhou and S. Wang, "Go Slow to Go Fast: Minimal On-Road Time Route Scheduling with Parking Facilities Using Historical Trajectory", in **VLDB Journal** (accepted).

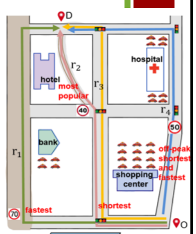
Data Sparsity

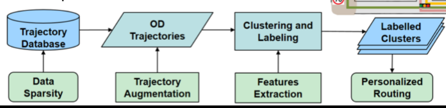
…no matter how much data you have, you don't have enough

+ Trajectory Clustering and Labeling

34

- Applications
  - Moving Behaviors Analysis
  - Personalized Routing
- Clustering
  - Data Sparsity: Origin-Destination Trajectory
  - Clusters Identification: Efficient Algorithms
- Labeling
  - Features: Fastest, Shortest, Most popular, Time-dependent



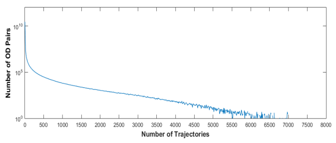


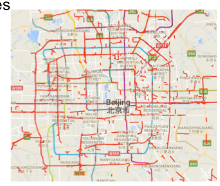
+ Trajectory Augmentation

35

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

No Robustness Validation Guarantee





[1]. Wang Yilun, Yu Zheng, and Yexiang Xue. "Travel time estimation of a path using sparse trajectories." ***SIGKDD*, *2014***.
[2]. Dai Jian, Bin Yang, Chenjuan Guo, and Zhiming Ding. "Personalized route recommendation using big trajectory data." ***ICDE, 2015***

D. He, B. Ruan, B. Zheng and X. Zhou, Origin-Destination Trajectory Diversity Analysis: Efficient Top-k Diversified Search, **MDM** 2018

Deep-Learning for Speed Predication

…many things we do is about using spatial data for prediction

Deep Learning + ITS = ?

To Deriving Knowledge from Data

+ CV = 

+ GO = 

+ ITS = ?

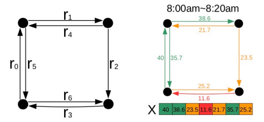
+ The Problem

38

- 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

$$X_t = [x_t^{r_0}, x_t^{r_1}, ..., x_t^{r_{|E|-1}}],$$

Problem: Given the historical observations $\{X_i | i = 1, ..., t\}$, this paper aims to predict $Y_t = \{X_j | j = t+1, ..., t+z\}$, where z is the number of time intervals to be predicted.

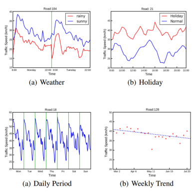


(a) A road network (b) A speed vector at 8:00am - 8:20am

+ Challenges

39

- 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

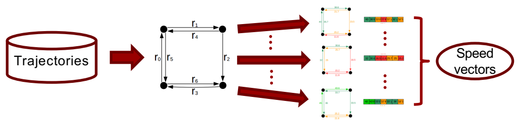


(a) Weather (b) Holiday

(a) Daily Period (b) Weekly Trend

+ Generating Speed Vectors

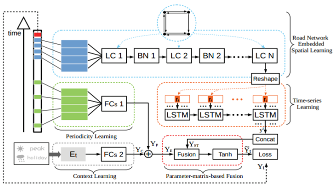
40



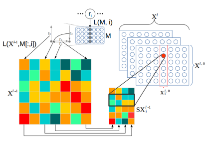
+ LC-RNN Model

41

- 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



LC-RNN model



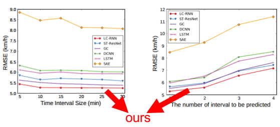
Look-up Convolution

+ Results

42

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

Methods	RMSE
SVR	10.245
H-ARIMA	11.867
SAE	8.471
LSTM	5.958
DCNN	6.085
GC	5.514
ST-ResNet	5.749
LC-3	5.437
LC-3-RNN	5.328
LC-3-RNN-E	5.296
LC-3-RNN-E-noFusion	5.392
LC-3-RNN-E-BN	**5.274**
LC-2-RNN-E-BN	5.319
LC-4-RNN-E-BN	5.285



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

Methods	RMSE
SVR	9.083
H-ARIMA	9.317
SAE	7.421
LSTM	5.274
DCNN	5.269
GC	4.921
ST-ResNet	5.087
LC-3-RNN-E-BN [ours]	**4.686**

Z. Lv, J. Xu, K. Zheng, P. Zhao, H. Yin and X. Zhou,, "LC-RNN: A Deep Learning Model for Traffic Speed Prediction", **IJCAI** 2018.

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

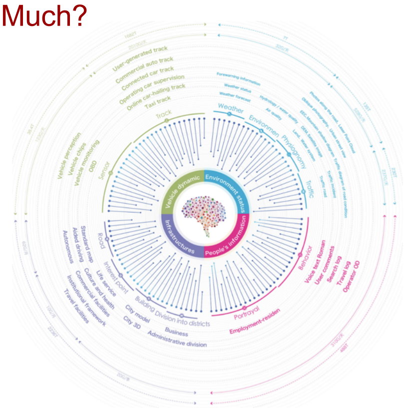
	Current	Daily
Company X (in-car navigation provider)	17.6TB	15M trajectories
Company Y (map app provider)	14.5TB	5M trajectories
Company Z (social network)	0.68TB	18M trajectories

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

+ NavInfo DataHIVE (minedata.cn)

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

+ How Much?



Trajectory Data Management System

…a common platform and API

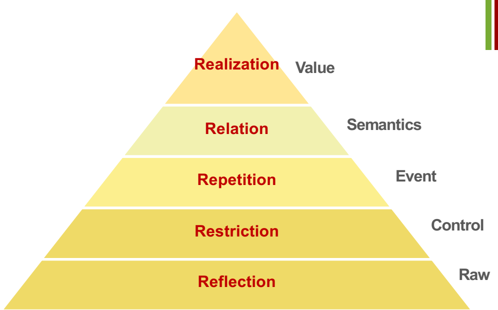
+ Why Common Platform?

50

- Universal
  - GPS, telecom tokens, social apps…
- Shared enterprise data
  - For monitoring, predication, 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…

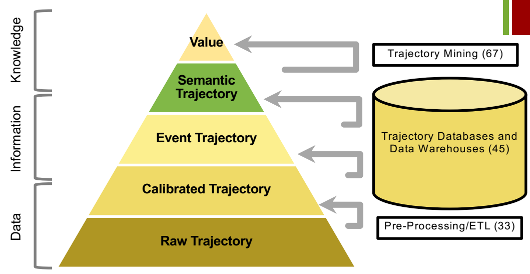
+ The 5R Approach

51



+ A Spatiotemporal Pyramid

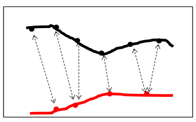
52



+ Trajectory Calibration

53

- 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?

H. Su, K. Zheng, H. Wang and X. Zhou, Calibrating Trajectory Data for Similarity-based Analysis, **SIGMOD** 2013

+ SparkDB

54

- 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. Wang, K. Zheng, X. Zhou and S. Sadiq, "SharkDB: An In-memory Column-oriented Trajectory Storage", **CIKM** 2014

Batch Fastest Path Queries

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

+ A Real Problem

56

- 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

57

- 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

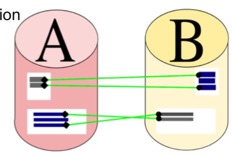
59

- 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?

60

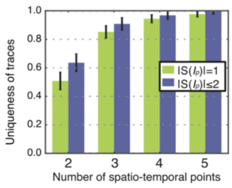
- Understanding the extent to which spatiotemporal data are distinctive is crucial to:
  - Location privacy protection
  - Entity resolution
  - Data integration
  - Spatial content retrieval
  - Personalized location recommendation
  - Driver performance assessment
  - …



+ Only 4 Points, Really?

61

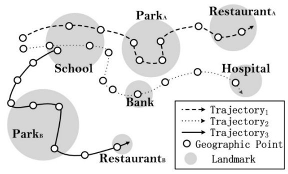
- "Unique in the Crowd: The privacy bounds of human mobility", Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen & Vincent D. Blondel, *Scientific Reports* (2013)
  - "four spatiotemporal points are enough to uniquely identify 95% of the individuals"



+ Everyone Has Mobility Signature?

62

- Spatial signature?
  - Regular: you visit frequently, such as your office building
  - Unique: you can be distinguished from others, like personal home address



+ Yes!

63

Top K	Precision
1	0.95987964
2	0.97241725
3	0.9779338
4	0.98194584
5	0.98445336
6	0.98545637
7	0.98696088
8	0.98746239
9	0.9884654
10	0.9889669

(1000 taxies)

- 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

Some Other New Research Problems

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

+ A List of Problems

65

- 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

66

- 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