Urban Data: Acquisition and Exploitation through Social and Crowd Sensing

Pierre Senellart

Futuring Cities, 15 May 2014

Collecting Urban Data

- Current approach:
  - Deploying sensor networks (air pollution, temperature, etc.)
  - Leveraging on other connected devices (car GPS systems, mobile cell towers, toll gates, etc.)
  - Personalized surveys
- Potential issues:
  - Cost of deploying new sensors
  - Data not publicly available, sometimes preciously guarded by companies as valuable assets
  - Privacy concerns, data may not be easily redistributed
  - Latency in installing new systems and acquiring new forms of data

The Web: An Abundant Source of Data

- Trillions of Web pages on amazingly diverse topics
- Hundreds of millions of messages posted each day on platforms such as Twitter
- Widespread use of social media in city inhabitants (68% of Singaporeans regularly use social media)
- User-contributed and open data of high quality (e.g., Open Street Map, RATP open data with all locations of bus stops)
- Crowdsourcing platforms (e.g., Amazon Mechanical Turk) can be used to obtain new data or to annotate existing data.
- A large part of this is fully accessible, free or cheap to obtain
- Give access to both objective and subjective information
- Complementary to traditional data collection
- The Social Web, the Crowd as virtual sensors

This talk

- 3 use cases for urban data collection from the Web (current research activities at Télécom ParisTech):
  - User Trajectories and Mobility Networks
  - Social Sensing for Trajectories of Moving Objects
  - Asking the Right Questions to the Crowd
- Outlook on general Web data management issues underlying these works

Outline

Semantics from users' GPS tracks

- Users often equipped with a smartphone with capability of recording regular timestamped locations from GPS data, WiFi network or cell tower triangulation
- Indeed, such information often already recorded to support some applications (Google Now, Frequent Locations, etc.)
- Tracks are noisy, incomplete, and not matched with actual transit networks
- Little added value for the user so far

Smarter urban mobility

- Problem: Smart and adaptive recommendations for mobility in cities (transit, bike rental, car, etc.)
- Challenge: Should take into account personal information (calendar, etc.), past trajectories, public information about transit networks and traffic
- Methodology: Map GPS tracks to routes of public transport, learn route patterns, infer destination while in transit and provide push suggestions

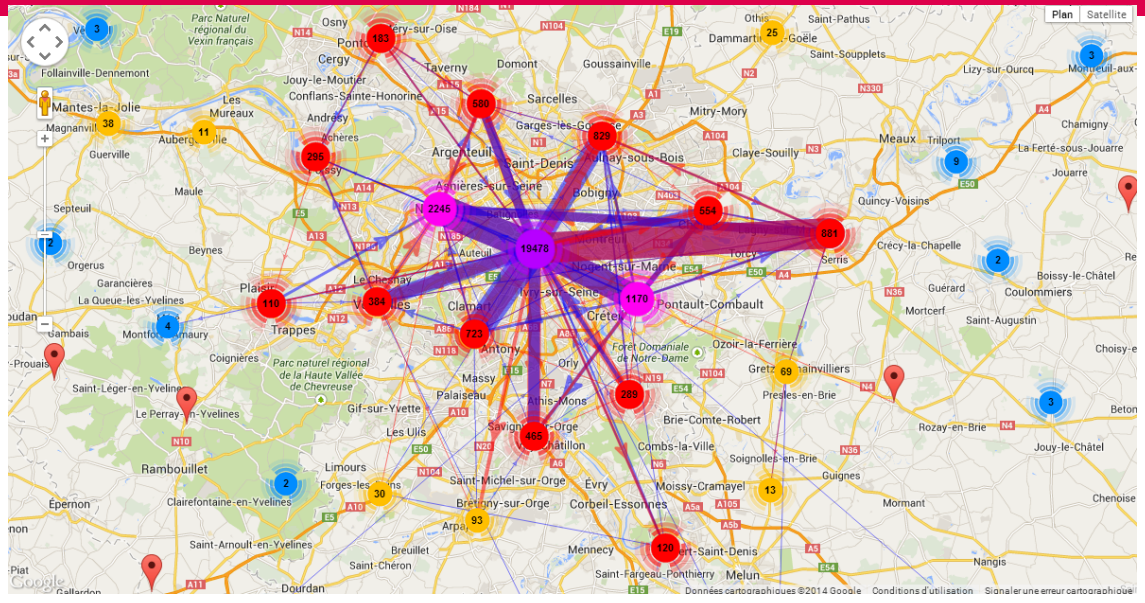
Outline

Geolocation from the Social Web

Much social Web content annotated with location information:

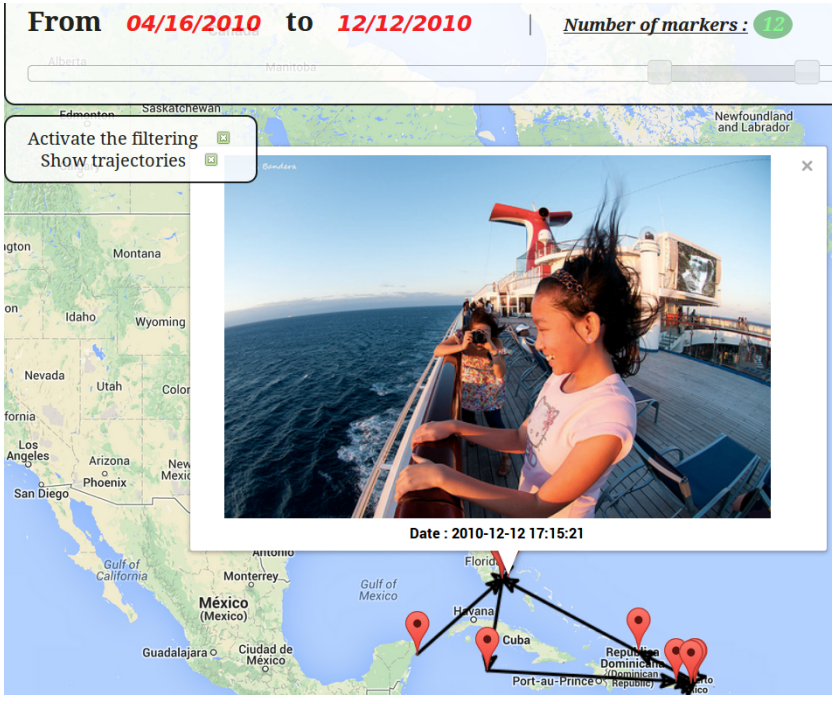
- Twitter or Facebook smartphone apps can add location to posts
- Pictures uploaded on Flickr or Instagram from GPS-equipped cameras (e.g., smartphones) often have geolocation data
- Possible to use information extraction techniques to extract location information from text and semi-structured information

Mobility in smart cities from social data



- Problem: Infer patterns of mobilities of populations within a city from social Web data (e.g., Instagram)
- Challenge: Partial and noisy data
- Methodology: Aggregate information from numerous users, to identify hotspots (e.g., tourist attractions) and paths between hotspots; cluster populations according to their mobility history

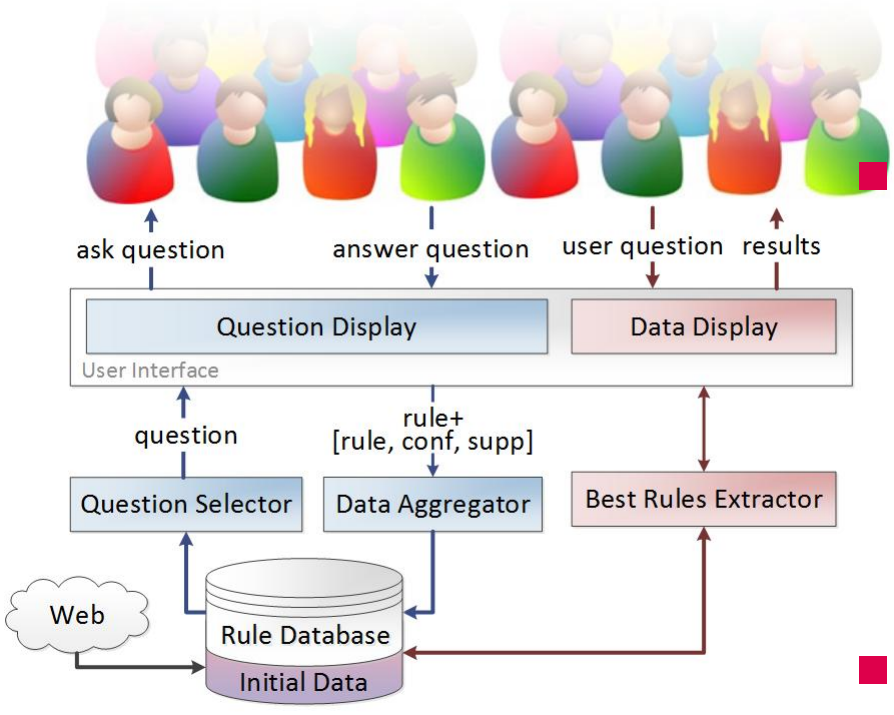
Social sensing of moving objects



- Problem: Infer trajectories and meta-information of moving objects from Web and social Web data (e.g., Flickr)
- Challenge: Uncertainty and inconsistency in extracted information
- Methodology: Data cleaning by filtering incorrect locations, and truth discovery to identify reliable sources

Outline

Optimizing crowd queries



- Problem: How to answer a specific mining need (e.g., mobility patterns) by asking a sequence of questions on crowdsourcing platforms
- Challenge: Crowd accesses are costly; individual questions asked (e.g., "How often do you use a bike?", "What is your most common transportation option to your workplace?") are not independent of each other
- Methodology: Maintain at all times a current knowledge of the world and estimate what is the best question(s) to ask given this knowledge.

Outline

Uncertain data is everywhere

Numerous sources of uncertain data:

- Measurement errors
- Data integration from contradicting sources
- Imprecise mappings between heterogeneous schemas
- Imprecise automatic processes (information extraction, natural language processing, etc.)
- Imperfect human judgment
- Lies, opinions, rumors

Structured data is everywhere

Data is structured, not flat:

- Variety of representation formats of data in the wild:
  - relational tables
  - trees, semi-structured documents
  - graphs, e.g., social networks or semantic graphs
  - data streams
  - complex views aggregating individual information
- Heterogeneous schemas
- Additional structural constraints: keys, inclusion dependencies

Intensional data is everywhere

Lots of data sources can be seen as intensional: accessing all the data in the source (in extension) is impossible or very costly, but it is possible to access the data through views, with some access constraints, associated with some access cost.

- Deep Web sources: Web forms, Web services
- The Web or social networks as partial graphs that can be expanded by crawling
- Outcome of complex automated processes: information extraction, natural language analysis, machine learning, ontology matching
- Crowd data: (very) partial views of the world
- Logical consequences of facts, costly to compute

Interactions between uncertainty, structure, intensionality

- If the data has complex structure, uncertain models should represent possible worlds over these structures (e.g., probability distributions over graph completions of a known subgraph in Web crawling).
- If the data is intensional, we can use uncertainty to represent prior distributions about what may happen if we access the data. Sometimes good enough to reach a decision without having to make the access!
- If the data is a semantic graph accessed by semantic Web services, each intensional data access will not give a single data point, but a complex subgraph.

State of the art and opportunities

Probabilistic databases cover limited structure variations, do not consider intensionality

Active and reinforcement learning deals with uncertainty and intensionality, but assumes trivial structures and simple goals

Crowdsourcing, focused crawling, deep Web crawling focus on specific applications of the uncertainty/structure/intensionality problem

Answering queries using views assumes simplistic cost models

Opportunities for Web data acquisition and exploitation systems that take all dimensions into account

Merci.

Contributions from, and joint work with:

Télécom ParisTech.	Talel Abdessalem, Antoine Amarilli, Mouhamadou Lamine Ba, Anis Bouriga, Sébastien Montenez
ENS Cachan & INRIA Saclay.	Serge Abiteboul, David Montoya
Tel Aviv University.	Yael Amsterdamer, Yael Grossman, Tova Milo
NUS.	Stéphane Bressan
A*STAR.	Huayu Wu