Predicting Crowds and Travel Times on Public Transportation

Marianne Winslet & Zhenjie Zhang
Advanced Digital Sciences Center & University of Illinois at Urbana-Champaign
ADSC is a University of Illinois research center in Singapore.
Why is ADSC in Singapore?

- Enhance Singapore's position as a hub for cutting-edge research in information technology.
- Provide Illinois with unique collaborative and funding opportunities for research with high potential impact.
What impact has ADSC had?  

| Best paper award nominations, other honors | 15 |
| Conference papers accepted last year | 87 |
| Journal papers accepted last year | 43 |

Scientific Impact

| ADSC technologies licensed | 5 |
| Startups facilitated in Singapore | 5 +1 |
| Patent, trade secret applications | 3 +1 |

Economic Impact (completed + underway)

| Interns | >100 |
| BS holders | >60 |
| PhD holders | >40 |

Workforce Development
Our group at ADSC does analytics.

Applications

Data Types, Bag of Tricks
- Short/long text, web data, spatiotemporal data, data streams, video
- Deep learning, topic modeling, queuing theory, differential privacy, game theory, ...

Systems
- Clouds, distributed systems, networking, DBMS internals, parallelization

Hardware
- Accelerators
Example recent projects

Applications
- Analyze the battery behavior on electric vehicles
- Optimize chiller plant energy efficiency

Data Types, Bag of Tricks
- Predict crowding, travel times in public transportation
- Human action detection from video
- Product text comment analysis (finished)
- Natural language database querying
- Real-time moving vehicle indexing and retrieval
- Privacy-preserving data analytics (finished)

Systems
- Resource allocation on cloud-based analytical system
- Fast failure recovery of distributed system
- Indexing incoming data stream at 1,000,000 tuples/sec
- 10+ patches to Apache Storm

Hardware
- Neural network in programmable hardware
- GPU-based real-time EEG analysis hardware
Mercury
Mercury: Current Progress

Prototype version is published. 

YouTube: https://www.youtube.com/watch?v=LataguUa4oQ

Data
- metro, bus, taxi, transportation card, weather, road condition

APP
- crowdedness prediction for metro, bus and taxi (iOS, Android supported)

System
- distributed streaming system (Spark, Storm)

Model
- deep learning model for prediction (recurrent neural network)

Commercialization
- location-based recommendation

Optimization
- improvement of model accuracy & user experience

Check the APP

24 Apr
Problem: Every morning

We have to face:
- crowded metro trains
- unavailable taxis
- slowly moving buses
- too many private cars

How can I avoid?
Real-time Crowdedness Prediction in 24 Hours

Solution: EARLY PREDICTION

Plan a day well

- availability of taxis
- driving time of buses
- number of passengers on trains
- early prediction

24 Apr. 2015
App can answer my question

Data Required: Real time + History

Real-time stream data for Short-term trend prediction

+ 

Historical big data for Deep learning model training
Subscribe the Prediction of Crowdedness

Metro line

Metro station

Prediction time

Metro map

Will be crowd at 8am?
Push the Notification of Crowdedness

Map of metro station

Number of passengers
- in station
- on platform
- on train

Trend in 1 hour

It's time to go!

Metro: Avoid crowdedness

*24 Apr. 2015*
**Metro: Data Behind**

**Raw Data**
- transportation card (only access records)
- metro operation data

**Route Planning**
- infer traveling path in-between any stations
- estimate traveling time (including waiting and interchange)

**Statistics**
- passengers in station (easy)
- passengers on platform (hard)
- passengers on train (very hard)

How to get the numbers?
Metro: Problems in Data

Ratio of Error-prone Data

- Passengers with incomplete access records: 6.85%
- Passengers in closed hours: 0.28%
- Passengers come in/out at the same station: 0.31%

So many problems!
Can I find a taxi?
Push the Notification of Availability

map of taxi stand

trend in 1 hour

I should wait for a while

prediction of taxi availability

FIND A TAXI

Taxi:
find a taxi

Push the Notification of Availability

I should wait for a while
**Raw Data**

- Taxi data (Only GPS, indication)

**Inference of Taxi Status**

- Build a city grid at granularity of 300m
- Recognize operating & off-service taxi
- Estimate velocity, heading and distance

**Statistics**

- # of passengers picked up in each grid
- # of passenger dropped off in each grid
- # of available taxis in each grid
- Average velocity in each grid

**Metro: Data behind**

How to locate a taxi?
Taxi: Problems in Data

- Status of available / busy taxi is incorrect
- Incorrect Status of off-service taxi
- Incorrect velocity
- Sparsity of available taxis in suburbs

Need data cleaning

24 Apr 2015
How can I go home?

Bus:

- Bus line: 955
- Direction: 济阳路三林路
- Target time: 06:00 PM
- Departure & destination:
  - 上海火车站
  - 恒丰路汉中路
  - 北京西路泰兴路
  - 陕西北路南京西路
  - 陕西南路永嘉路
  - 打浦桥
  - 徐家汇路黄陂南路
  - 斜桥
  - 陆家浜路大兴街
  - 陆家浜路跨龙路
It's too slow!

Bus: Manage my time

Push the Notification of Driving Time

driving time between departure and destination

trend in 1 hour

bus route
Bus: Data Behind

- **Raw Data**
  - Bus data (records of arriving & leaving stop)

- **Inference of Bus Operation**
  - Correct bus route by arriving time
  - Calculate driving time in-between stops
  - Correct the records of skipping stops

- **Statistics**
  - Compute driving time during off-peak hour
  - Compute driving time during peak hour

How to know the traveling time?
Bus: Problems in data

- Skipped stops by the buses
- Incomplete bus arriving records
- Incorrect direction of bus operation
- Encoding error of Chinese
Real-time Onsite Recommendation

Based on user’s subscriptions and current traffic conditions, the onsite recommendation system provides real-time recommendations. For example, if you have a cup of coffee first, it might be more convenient to take the subway to your destination. The advertisement is information that can be used to make better decisions about your commute.

Commercialization: LB Recommendation

24 Apr
A Solution for City Management

reward passengers who travel during off-peak hours

Rewards: staggered shifts

I can earn rewards
**Dataset for Training**
1 Apr. 2015 – 23 Apr. 2015

**System:**
**Data Mining**

- **Data Cleaning**
  - pre-process massive data on Hadoop

- **Model Training**
  - train deep learning model offline on Spark

- **Real-time Predicting**
  - compute near-future trend online via Storm

**Diagram:**
- **Notification to App**
- **Output trained model**
- **Online updating**
- **Prepare training data**
- **Archive historical data**
- **Deep Learning Model Training**
- **Data Cleaning**
- **Data Integration**
- **Real-time Analysis**
- **Storm**
- **Spark & Hadoop**
- **Raw data**

**How to train a model?**
System: Real-time prediction

- **Real-time**
  - Response time < 1ms

- **High Concurrency**
  - 150K times prediction per second per node

- **Elastic Computation**
  - Increase / decrease nodes on demand

- **Highly Usable**
  - Distributed fault tolerance

Dataset for Testing
2015.4.24 - 2015.4.30
Integration of Road and Weather Information to Improve Model Accuracy

- weather
- accident on road
- traffic index of city road
- closure of ramp

MODEL: MULTIPLE FACTORS

Consider other factors?
Efficient Business

Comfortable Travel

Smart Transportation

Safe Driving

Green City

VALUABLE

A new day coming

24 Apr 2015
Problem Analysis

**Problem**
- Crowd Information Collection
- Staff Management
- Guest Engagement

**Cause**
- Lack of Effective Perception
- Lack of Prediction Capability
- Lack of Contact Channel and Personalized Recommendation

**Technical Intuition**
- Real-time Video Analytics
- Predictive Analysis by integrating data
- Targeted Incentives on smart phones

---

Using Queue Marker as Proxy for Ridership Information:
- Duty personnel manually observes which queue marker the last person in the queue is at via CCTV.
- Waiting time is estimated by the last queue marker's position from the average waiting time.
- Data is manually entered into an Excel file and analyzed for average waiting time.
- Cycle is repeated every 10 minutes.

Unable to Capture Tram & Bus Ridership Information:
- Two earlier considerations made were not workable:
  - People counting mats: guests cannot see or access them.
  - RFID tags/wearable: no practice exists for ridership information.
Overall Workflow

- Retrieve Video Feeds
- Crowd Evaluation
- Crowd Prediction
- Monitoring → Integration → Prediction
- Staff Notification
- Track Staff Locations
- Viewing
- Response
- Check Bus Waiting Time
- Alternative Transportation
- Checking
- Suggestion
- Redeem F&B Voucher/Coupon

- 300 guests at bus stop
**Predictive Analytics**
- Video Analysis
- Density Evaluation
- Density Prediction

**Staff Engagement**
- Video Sharing
- Location Tracking

**Guest Engagement**
- Coupon Recommend
- Coupon Redeem

**Predictive Analytical System**
- Video Labs
- Mercury

**Engagement System**
- IMS
- Recommender
- Video Sharing

**System Architecture**
Data Cleaning

- pre-process massive data on Hadoop

Model Training

- train deep learning model offline on Spark

Real-time Predicting

- compute near-future trend online via Storm

Use Cases:

Real-Time Traffic Prediction
Sustainable Business Model

Redeem

Coupon

Recommender

Rebate
Economic Analysis

One-Time Hardware Investment

- Server for video processing, predictive analysis and app hosting
- Software on analytics and mobile apps

Operational Cost

- Electricity on computation
- Software maintenance

Incomings

- Each coupon charges the vendor dozens of cents