User Clustering with GPS Trajectories

Yan Wang

DVMM Lab, Electrical Engineering Department

Dec. 8, 2011
Overview

▶ Introduction
  ▶ Problem definition
  ▶ Dataset
▶ Approach
  ▶ Visualization and analysis
  ▶ Preprocessing
  ▶ Feature design
  ▶ Visualization and analysis
  ▶ Refined feature design
  ▶ Clustering approaches
▶ Experiments
  ▶ Abstract indicators
  ▶ Visualization and analysis
▶ Conclusion
  ▶ What I did
  ▶ What I learned
Problem definition

- User clustering with GPS trajectories
Problem definition

- User clustering with GPS trajectories
- It’s important
  - Spatial-temporal data: general
  - Applications
- It’s challenging
  - Spatial-temporal data: complex
  - Geographical data: rich context
Dataset

- GeoLife from Microsoft Research Asia
  - Latitude, Longitude, Timestamp, UserID
  - Large-scale
    - 167 users, 2 years, 17K+ trajectories, 1M+ km, 48K hours
- Subset in urban area of Beijing
- Why choosing GeoLife?
  - Suitable for the problem
  - Large enough
Approach

▸ Framework
  ▸ Preprocessing
  ▸ Feature extraction
  ▸ Clustering

▸ Why?
Visualization and analysis

▶ What does the overall data look like?

▶ Problems?
  ▶ Not informative enough!
  ▶ More interested in *intentionally* visit
Preprocessing

- Filter high-speed points

Solution!

- Noise filtered
- Enough points preserved (4M+)
Visualization and analysis

- What does one user look like?

- Problems?
  - Outliers!
Preprocessing

- Discard points inconsistent with local average

- Solution!
  - Smooth and reasonable trajectories
Feature extraction

- Intuition
  - Traveling time
    - Morning? Afternoon? Evening?
    - Hour in timestamp
  - Traveling area
    - Northwest? Central?
    - 1 from 100 regions in Beijing
  - Category
    - Restaurants? Bookstores? Shopping malls? Parks?
Visualization and analysis

- Traveling time
  - Good
    - Reasonable
    - Balanced
Visualization and analysis

- Traveling area
  - Right figure
- Very unbalanced
  - Why?
- Periodical
  - Why?
  - Problem of distance
Visualization and analysis

- Traveling area
  - Right figure
- Very unbalanced
  - Why?
- Periodical
  - Why?
  - Problem of distance
Refined feature design

- 1D ⇒ 2D

- Solution!
Incorporate geographical contexts

- A sample scenario
- What we need
  - Describe the “category” of a place
- What we have
  - Google Place API
  - (Lat, Lon, Type, Search range) \(\Rightarrow\) Detailed info about local business
- What to do
  - # Restaurants, # Bookstores, # Shopping malls, # Parks
Visualization and analysis

- Geographical distribution

- Homogeneous
  - Why?
  - Google Place API returns 20 results at most.
Refined feature design

- Decrease search range

Solution!
  - More discriminative
Refined feature design

- Feature
  - Mean of traveling time
  - Mean of traveling area
  - Mean of “category” feature
- Build Bag-of-Words feature
- Normalization
  - 0-mean, 1-standard-deviation
Visualization and analysis

- Problems?
  - Missing values
- Solution
  - Fill them with mean
Clustering

- Approaches
  - K-Means
  - Spectral Clustering
  - Affinity Propagation: no cluster # required ahead
- Feature selection with PCA
Experiments

- KMeans
- Calinski Index
  - $\frac{\text{Between Cluster SS}}{(K-1)}$ / $\frac{\text{Within Cluster SS}}{(N-K)}$
  - The larger, the better
- No need to normalize
Visualization and analysis

- $k = 6$, PCA + KMeans

- KMeans cannot give stable results

- The data is not really suitable for clustering
Visualization and analysis

- Affinity Propagation thinks there should be 5 clusters
- It can provide a stable result
Visualization and analysis

- Spectral Clustering provide non-sense results: one huge cluster with tiny “outlier” clusters.

- Actually a good method!
Conclusion

- What I did
  - Iterative feature design
    - For spatial-temporal data
    - Incorporate rich context of geographical data
  - Feature selection
    - Improve clustering performance
  - Clustering approaches comparison

- What I learned
  - *LOOK INTO THE DATA! IN EVERY STEP!*
    - Discover and solve problem efficiently
    - Solid practice
  - Ask why
    - Dig internal insights
    - Detect bugs
Thank you

- Slides and more demos available at
  - http://lab.grapeot.me/
- Discuss and contact
  - https://www.facebook.com/grapeot/