TOPTRAC: Topical Trajectory Pattern Mining

Source: KDD 2015

Advisor: Jia-Ling Koh

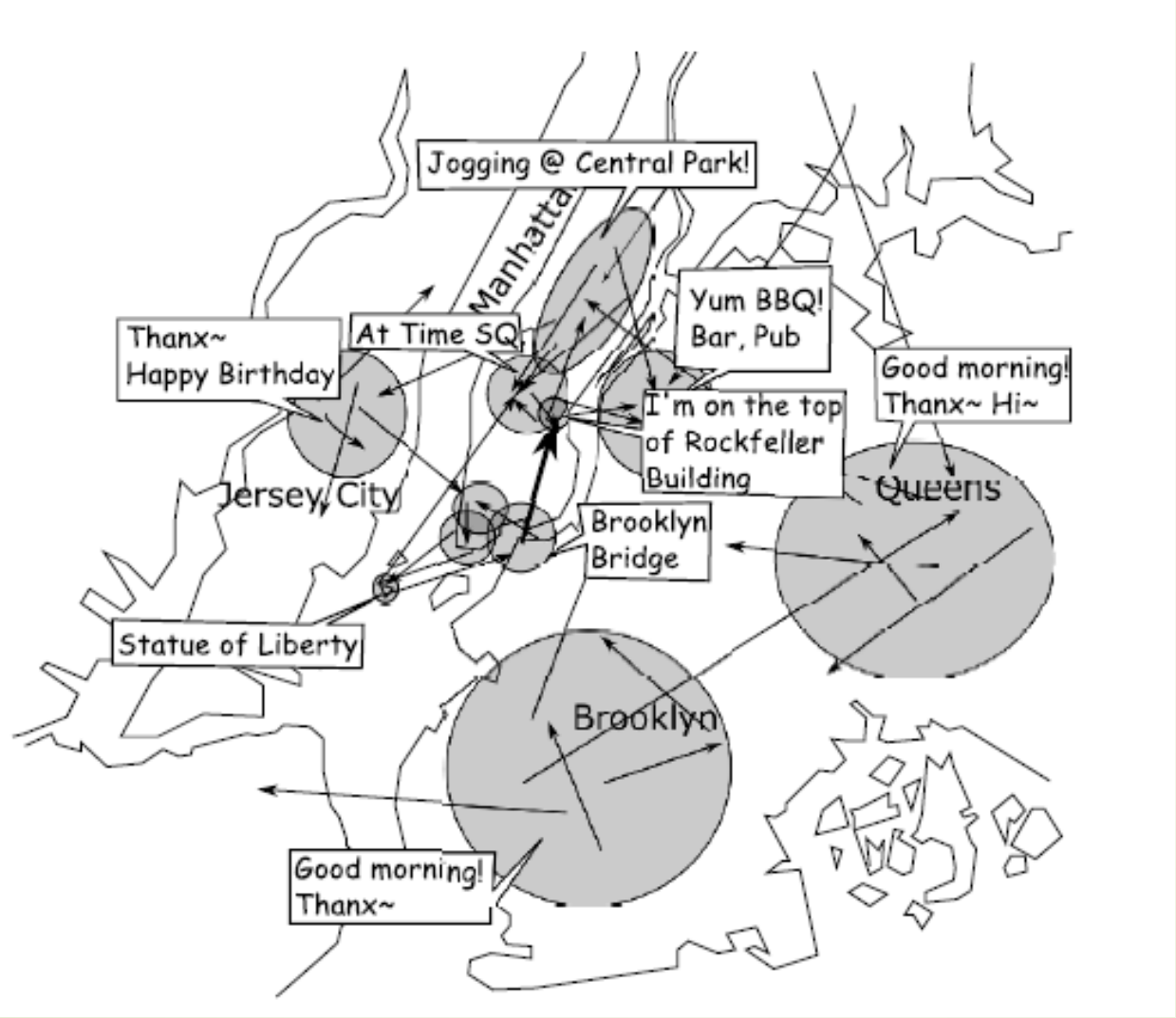
Speaker: Hsiu-Yi,Chu

Date: 2018/1/21

Outline

- **Introduction**
- Method
- Experience
- conclusion

Introduction



Introduction

- Goal

  Topical trajectory mining problem:

  Given a collection of geo-tagged message trajectories, it's to find **topical transition pattern** and the **top-k transition snippets** which best represent each transition pattern

Introduction

	(message, location)
s_1	$m_{1,1}$=(Start today's tour, L_2) $m_{1,2}$=(Time square, L_1)
s_2	$m_{2,1}$=(Apple store, L_5) $m_{2,2}$=(Running in Central park, L_3)
s_3	$m_{3,1}$=(Many people @ Time SQ, L_1) $m_{3,2}$=(Central park zoo, L_3)
s_4	$m_{4,1}$=(Liberty Statue tour, L_2) $m_{4,2}$=(I'm at Time square, L_1)
s_5	$m_{5,1}$=(Now, MoMA, L_4) $m_{5,2}$=(Metro. museum of art, L_3)
s_6	$m_{6,1}$=(At MoMA, L_4) $m_{6,2}$=(Museum of natural history, L_3)
	...

- Transition pattern:

  "Statue of Liberty" "Time Square"

- Transition snippet:

  ($m_{1,1}$, $m_{1,2}$)in s_1

  ($m_{4,1}$, $m_{4,2}$)in s_2

Introduction

- Definition
  - Trajectory(s_t)
    - geo-tagged message ($m_{t,i}$)

      Geo-tag $G_{t,i}$: 2-dim vector($G_{t,i,x}$,$G_{t,i,y}$)

      Bag-of-word $w_{t,i}$: N words{$w_{t,i,1}$,…, $w_{t,i,n}$}

Introduction

- Definition
  - Latent semantic region:

    a geographical location where messages are posted with the same topic preference
  - Topical transition pattern:

    a movement from one semantic region to another frequently

Outline

- Introduction
- **Method**
- Experience
- conclusion

Method

- Generative Model
  - Assume there are **M** latent semantic regions
    K hidden topics
    in the collection of geo-tagged messages

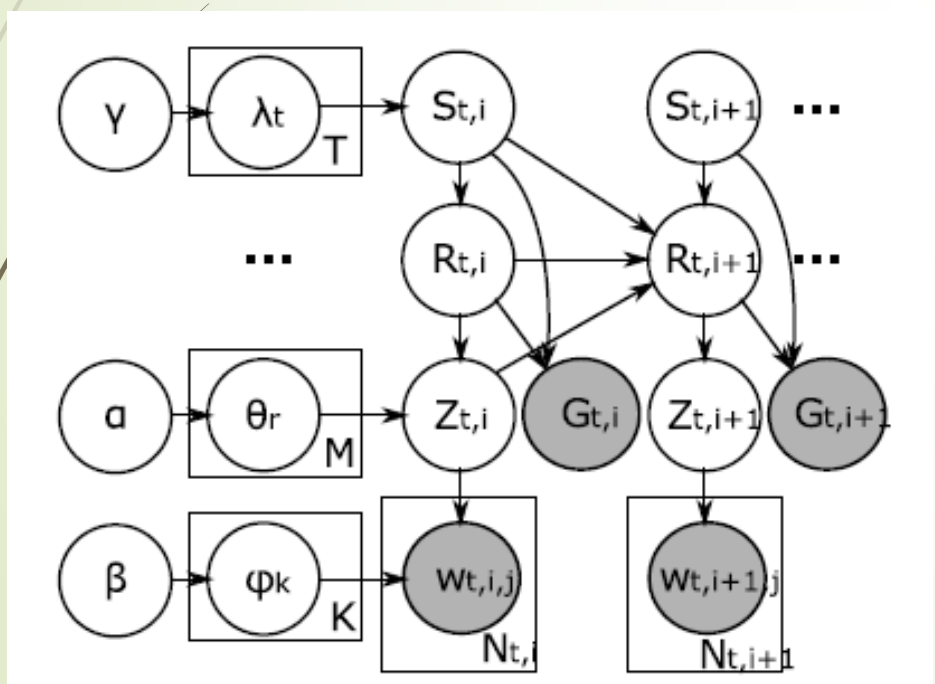
Method

- Variables

	Description	Approx. vars.
$\vec{\theta}_r$	Topic distribution in region r	$q(\vec{\theta}_r)$=Dir($\vec{a}_r$)
$\vec{\phi}_k$	Word distribution for the k-th topic	$q(\vec{\phi}_k)$=Dir($\vec{b}_k$)
$\vec{\lambda}_t$	Bernoulli distribution for the relationship to the local context in s_t	$q(\vec{\lambda}_t)$=Beta($\vec{c}_t$)
$S_{t,i}$	Random variable representing whether $m_{t,i}$ is in the local context or not	$q(S_{t,i})$ $=\langle\sigma_{t,i,0}, \sigma_{t,i,1}\rangle$
$R_{t,i}$	Random variable for the latent semantic region of $m_{t,i}$	$q(R_{t,i})$ $=\langle\rho_{t,i,1}, .., \rho_{t,i,M}\rangle$
$Z_{t,i}$	Random variable that indicates the latent topic used to generate $m_{t,i}$	$q(Z_{t,i})$ $=\langle\zeta_{t,i,1}, .., \zeta_{t,i,K}\rangle$
$\vec{\delta}$	Probability distribution of selecting a starting latent semantic region	-
$\vec{\delta}_{r,k}$	Transition probability distribution from region r with topic k	-
$\vec{\mu}_r$, Σ_r	Mean and covariance matrix of r	-
$\vec{\alpha}$,$\vec{\beta}$,$\vec{\gamma}$	Hyper-parameters for the Dirichlet priors of $\vec{\theta}_r$, $\vec{\phi}_k$ and $\vec{\lambda}_t$ respectively	-

Method

- Generative process



For each region $r=1,\ldots,M$:

- Select a categorical distribution: $\vec{\theta}_r \sim \text{Dir}(\vec{\alpha})$

For each topic $k=1,\ldots,K$:

- Select a categorical distribution: $\vec{\phi}_k \sim \text{Dir}(\vec{\beta})$

For each sequence $s_t=\langle m_{t,1},\ldots,m_{t,N_t}\rangle \in \mathbb{C}$:

- Select a Bernoulli distribution: $\lambda_t \sim \text{Beta}(\vec{\gamma})$
- For each message $m_{t,i} = (\vec{G}_{t,i}, \mathbf{w}_{t,i})$:
  - Decide the status of $m_{t,i}$: $S_{t,i} \sim \text{Bernoulli}(\lambda_t)$
  - If $(S_{t,i} = 0)$
    - Select a region: $R_{t,i} \sim \text{Uniform}(1/M)$
    - Generate a geo-tag: $\vec{G}_{t,i} \sim \text{Uniform}(f_0)$
  - If $(i{=}1 \wedge S_{t,1}{=}1) \vee (i{\geq}2 \wedge S_{t,i-1}{=}0 \wedge S_{t,i}{=}1)$
    - Select a region: $R_{t,i} \sim \text{Categorical}(\vec{\delta}_0)$
    - Generate a geo-tag: $\vec{G}_{t,i} \sim N(\mu_{R_{t,i}}, \Sigma_{R_{t,i}})$
  - Else (i.e., $i{\geq}2 \wedge S_{t,i-1} = 1 \wedge S_{t,i} {=}1$)
    - Select a region: $R_{t,i} \sim \text{Categorical}(\vec{\delta}_{R_{t,i-1}, Z_{t,i-1}})$
    - Generate a geo-tag: $\vec{G}_{t,i} \sim N(\mu_{R_{t,i}}, \Sigma_{R_{t,i}})$
  - Select a topic: $Z_{t,i} \sim \text{Categorical}(\vec{\theta}_{R_{t,i}})$
  - Generate a message: $\mathbf{w}_{t,i} \sim \text{Multinomial}(\vec{\phi}_{Z_{t,i}})$

Method

- Select Geo-tag $G_{t,i}$ according to a 2-dimensional Gaussian probability function:

$$f_{R_{t,i}}(\vec{G}_{t,i}) = \frac{1}{2\pi\sqrt{|\Sigma_{R_{t,i}}|}} \exp\left(-\frac{1}{2}(\vec{G}_{t,i} - \vec{\mu}_{R_{t,i}})^{\top}\Sigma_{R_{t,i}}^{-1}(\vec{G}_{t,i} - \vec{\mu}_{R_{t,i}})\right)$$

Method

- Likelihood

$$\mathbb{L} = \prod_{k=1}^{K} \int_{\vec{\phi}_k} Dir(\vec{\phi}_k; \vec{\beta}) \prod_{r=1}^{M} \int_{\vec{\theta}_r} Dir(\vec{\theta}_r; \vec{\alpha}) \prod_{t=1}^{t} \int_{\vec{\lambda}_t} Beta(\vec{\lambda}_t; \vec{\gamma}) Pr_{\Omega}(s_t) d\vec{\lambda}_t d\vec{\theta}_r \vec{\phi}_k$$

$$Pr(m_i|R_i, Z_i) = f_{R_i}(\vec{G}_i) \cdot \prod_{j=1}^{N_i} \phi_{Z_i, w_j},$$

$$Pr_{\Omega}(\langle m_{1:N} \rangle) = \sum_{R_{1:N}, S_{1:N}, Z_{1:N}} \lambda_{S_1} \cdot Pr(R_1|S_1) \cdot \theta_{R_1, Z_1} \cdot Pr(m_1|R_1, Z_1) \cdot \prod_{i=2}^{N} \lambda_{S_i} \cdot Pr(R_i|S_i, S_{i-1}, R_{i-1}, Z_{i-1}) \cdot \theta_{R_i, Z_i} \cdot Pr(m_i|R_i, Z_i)$$

$$Pr(R_i|S_i, S_{i-1}, R_{i-1}, Z_{i-1}) = \begin{cases} 1/M & \text{if } S_i=0, \\ \bar{\delta}_{R_i} & \text{if } (i=1 \text{ and } S_i=1) \text{ or } (i \geq 2 \text{ and } S_{i-1}=0), \\ \delta_{R_{i-1}, Z_{i-1}, R_i} & \text{if } i \geq 2 \text{ and } S_i=1 \text{ and } S_{i-1}=1. \end{cases}$$

Method

- Variational EM Algorithm
  - Maximum likelihood estimation

$q(\vec{\theta}_r)=\mathrm{Dir}(\vec{a}_r)$
$q(\vec{\phi}_k)=\mathrm{Dir}(\vec{b}_k)$

$q(\vec{\lambda}_t)=\mathrm{Beta}(\vec{c}_t)$

$q(S_{t,i})$
$=\langle \sigma_{t,i,0}, \sigma_{t,i,1} \rangle$

$q(R_{t,i})$
$=\langle \rho_{t,i,1}, .., \rho_{t,i,M} \rangle$
$q(Z_{t,i})$
$=\langle \zeta_{t,i,1}, .., \zeta_{t,i,K} \rangle$

$\vec{\mu}_r$, Σ_r

Method

- Finding the Most Likely Sequence
  - Notations:

- $s_t[i]$: the subsequence of s_t which starts at the first message and ends at the i-th message of s_t.
- $\bar{\pi}[i]$: the maximum probability to generate $s_t[i]$ when $m_{t,i}$ is submitted without any local context (i.e., $S_{t,i}=0$).
- $\pi[i, r, k]$: the maximum probability to create $s_t[i]$ when $m_{t,i}$ has the local context, its latent semantic region is r and the latent topic is k (i.e., $S_{t,i}=1 \wedge R_{t,i}=r \wedge Z_{t,i}=k$).
- $\Pi[i]$: the maximum probability to generate $s_t[i]$ which is computed as $\max\{\bar{\pi}[i], \max_{1\leq r\leq M, 1\leq k\leq K} \pi[i, r, k]\}$.

Method

- Compute $\bar{\pi}[i]$:
  - $$\bar{\pi}[i] = \Pi[i-1] \cdot \max_{1 \le r \le M, 1 \le k \le K} \lambda_0 \frac{1}{M} \theta_{r,k} Pr(m_{t,i} | R_{t,i} = r, Z_{t,i} = k).$$
- Compute $\pi[i, r, k]$:
  - case1: $S_{t,i-1} = 0$; case2 : $S_{t,i-1} = 1$
  - $$\pi[i, r, k] = \max \left\{ \begin{array}{ll} \Pi[i-1] \cdot \lambda_1 \cdot \delta_{0,r} \cdot \theta_{r,k} \cdot Pr(m_{t,i} | R_{t,i} = r, Z_{t,i} = k), & \text{Case (1)} \\ \max_{1 \le r' \le M, 1 \le k' \le K} \{\pi[i-1, r', k'] \cdot \lambda_1 \cdot \delta_{r',k',r} \cdot \theta_{r,k} & \\ \cdot Pr(m_{t,i} | R_{t,i} = r, Z_{t,i} = k)\} & \text{Case (2)} \end{array} \right\}$$

Method

- Finding Frequent Transition Patterns
  - $s_t' = \{(s_{t,1}, r_{t,1}, z_{t,1}), \ldots, (s_{t,n}, r_{t,n}, z_{t,n})\}$
  - Transition Patterns = $\{(r_1, z_1)(r_2, z_2)\}$
  - Start with $(1, r_1, z_1)$ and ends with $(1, r_2, z_2)$
  - τ : minimum support

Method

- Example
  - s_1'={(0,1,1)(1,1,2)(1,2,1)}, s_2'={(1,1,2)(0,2,1)(1,2,1)} with τ = 2 → {(1,2)(2,1)} is a transition pattern

- Top-k transition snippets
  - k largest probabilities of

$$\delta_{r_1,z_1,r_2} Pr(m_{t,i}|R_{t,i} = r_1, Z_{t,i} = z_1) Pr(m_{t,j}|R_{t,j} = r_2, Z_{t,j} = z_2),$$

Outline

- Introduction
- Method
- **Experience**
- conclusion

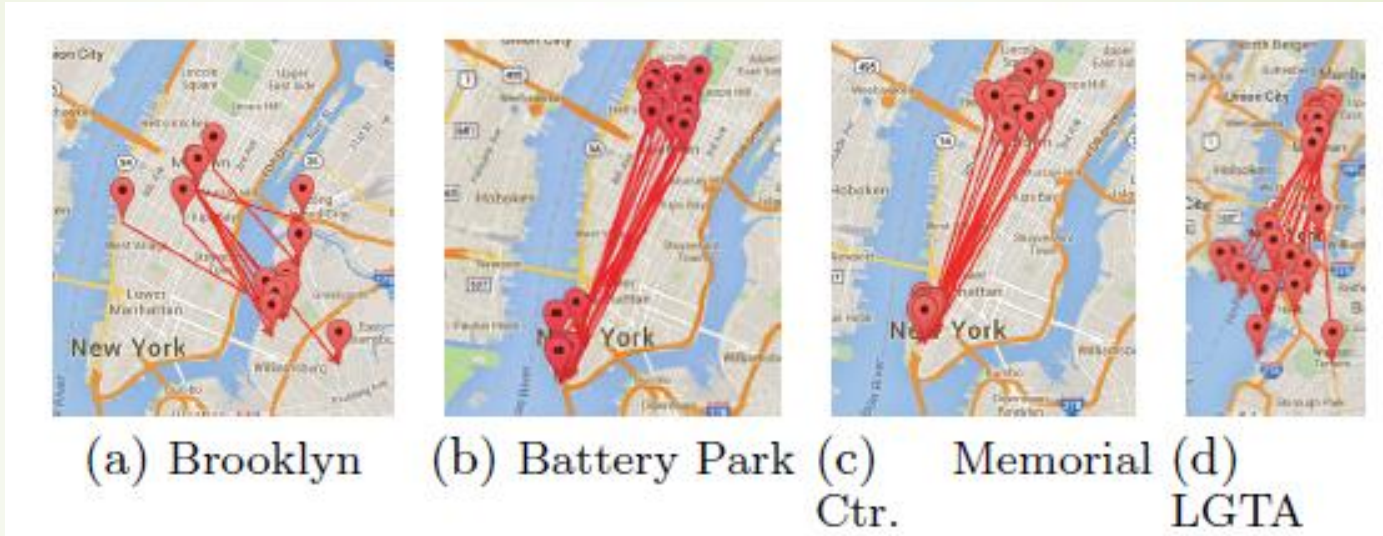
Experience

- Data sets
  - NYC
    - 9070 trajectories, 266808 geo-tagged messages
    - M = 30, K = 30, τ = 100
  - SANF
    - 809 trajectories,19664 geo-tagged messages
    - M = 20, K = 20, τ = 10

Experience

- Baseline
  - LGTA
    - Run the inference algorithm and find frequent trajectory patterns similar in page15,16
  - NAÏVE
    - First groups messages using EM clustering
    - Cluster the messages in each group with LDA

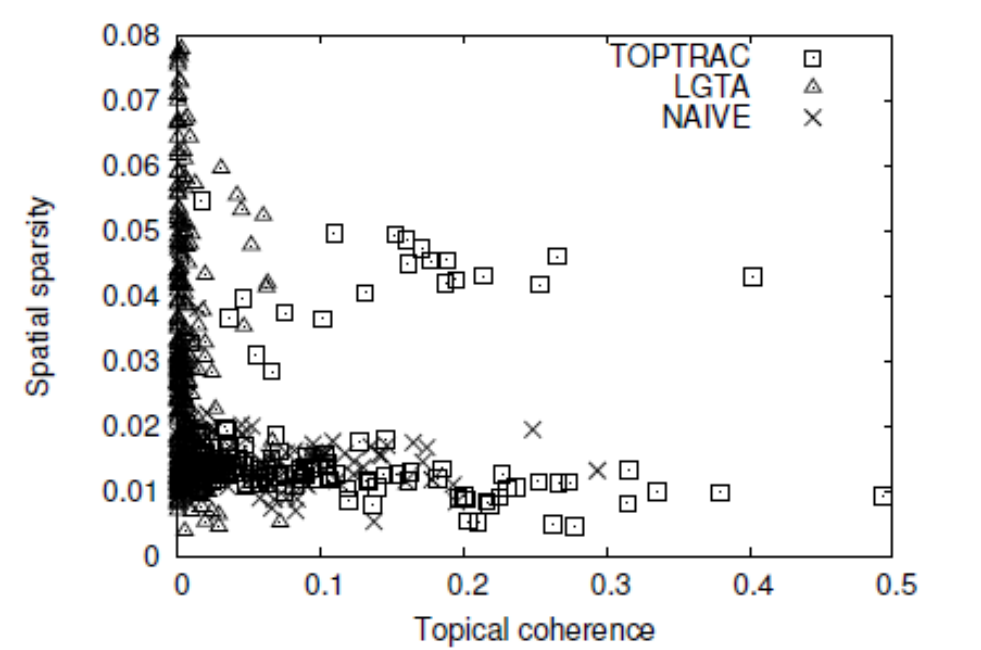
Experience



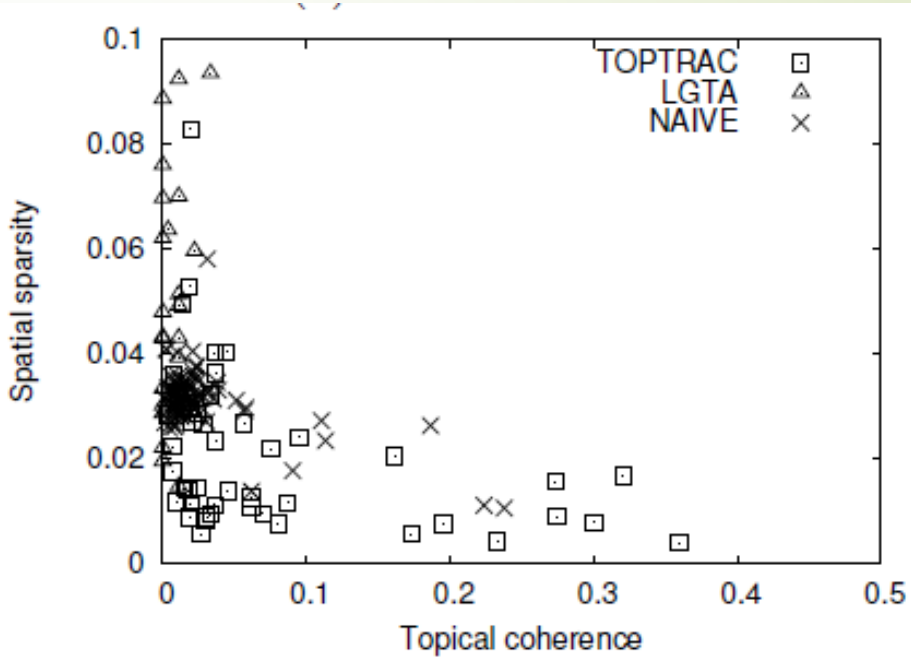
(a) Brooklyn (b) Battery Park (c) Memorial Ctr. (d) LGTA

(e) Sausalito (f) Chinatown (g) Winery (h) NAIVE

Experience

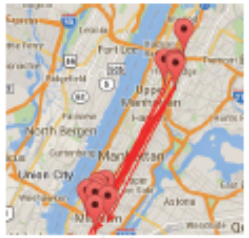


(a) NYC data set



(b) SAF data set

Experience



(a)

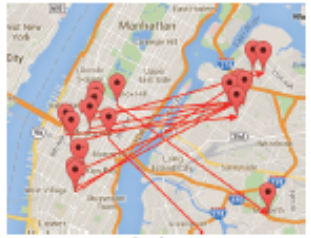
(1) @ Yankee Stadium http://t.co/*
(2) This was for dinner. #NYC #bestrestaurant #foodporn @ The Halal Guys http://t.co/*
(1) I'm at Yankee Stadium - @mlb (Bronx, NY) http://t.co/*
(2) I'm at Wa Bar + Kitchen - @wabarnyc (New York, NY) http://t.co/*
(1) We are here for the gamew (@ Yankee Stadium Gate 8 w/ 7 others) http://t.co/*
(2) I'm at O'Donoghues Bar & Restaurant - @odonoghuests (New York, NY) w/ 2 others http://t.co/*

(b) From 'Yankee Stadium' to Downtown for 'Dinner'

(c)

(1) Morning @ Central Park 5av and 59th Street NYC http://t.co/*
(2) Just posted a photo @ Egg Restaurant http://t.co/*
(1) I'm at The Biergarten at The @StandardHotels (New York, NY) w/ 17 others http://t.co/*
(2) Brunching (at @Allswell_nyc) http://t.co/*
(1) #nyc icon. @ Empire State Building http://t.co/*
(2) Brunch time. #nyc #brooklyn @ Juliette Restaurant http://t.co/*

(d) From Downtown to Brooklyn for 'Brunch'

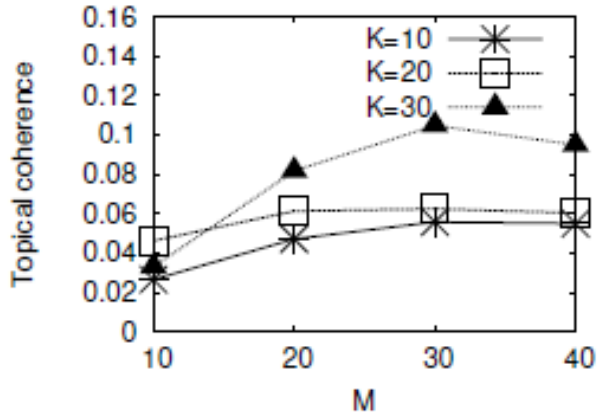


(e)

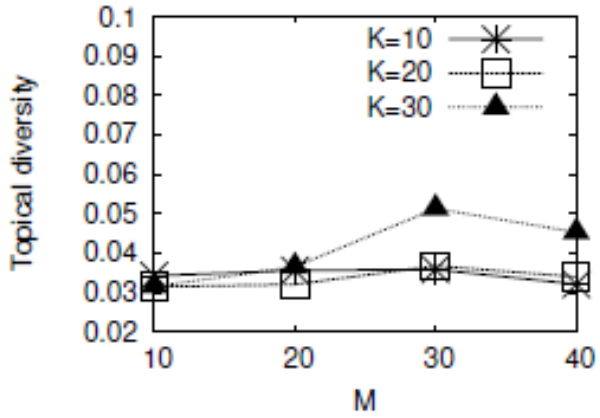
(1) I'm at Apple Store (New York, NY) w/ 7 others http://t.co/*
(2) I'm at @DunkinDonuts (Maspeth, NY) https://t.co/*
(1) I'm at Chipotle Mexican Grill (New York, NY) w/ 4 others http://t.co/*
(2) I'm at Brooklyn Bagel & Coffee Company (Astoria, NY) w/ 2 others http://t.co/*
(1) I'm at Union Square Park - @nycparks (New York, NY) w/ 31 others http://t.co/*
(2) I'm at New York City Bagel & Coffee House - @nycbch (Astoria, NY) http://t.co/*

(f) From Downtown to Brooklyn for 'Bagel'

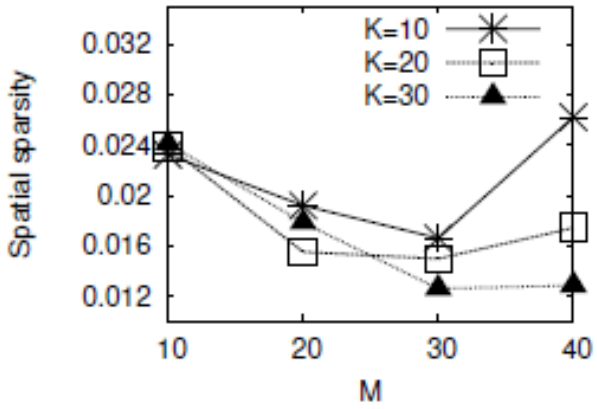
Experience



(a) Topical coherence



(b) Topical anti-diversity



(c) Spatial sparsity

Experience

TOPTRAC					LGTA					NAIVE				
M	Q1	Q2	Q3	Q4	*M*	Q1	Q2	Q3	Q4	*M*	Q1	Q2	Q3	Q4
20	0.0816	0.0178	0.0576	0.0364	70	0.0048	0.023	0.0602	0.0068	70	0.042	0.0142	0.0326	0.0133
30	0.1047	0.0126	0.0459	0.0512	100	0.006	0.0301	0.0595	0.0067	100	0.061	0.0138	0.038	0.022
40	0.102	0.0124	0.0463	0.0548	150	0.0061	0.025	0.05	0.008	150	0.057	0.013	0.0293	0.0169

(a) NYC data set

TOPTRAC					LGTA					NAIVE				
M	Q1	Q2	Q3	Q4	*M*	Q1	Q2	Q3	Q4	*M*	Q1	Q2	Q3	Q4
10	0.0867	0.0293	0.0941	0.0396	30	0.0272	0.7340	0.8454	0.0134	50	0.0072	0.6274	0.678	0.015
20	0.0772	0.0196	0.0828	0.0464	40	0.0277	0.4553	0.4928	0.0125	70	0.0414	0.0261	0.0707	0.0487
30	0.0915	0.0172	0.1018	0.0501	50	0.0072	0.3274	0.6780	0.015	100	0.0256	0.0311	0.0581	0.0522

(b) SANF data set

Figure 10: Quality of clusters (Q1:topical coherence, Q2: spatial sparsity, Q3: distance, Q4: topical anti-diversity)

Outline

- Introduction
- Method
- Experience
- **conclusion**

Conclusion

- Propose a trajectory pattern mining algorithm, called TOPTRAC, using probabilistic model to capture the spatial and topical patterns of users.
- Developed an efficient inference algorithm for our model and also devised algorithms to find frequent transition patterns as well as the best representative snippets of each pattern.