Who, Where, When and What: Discover Spatio-Temporal Topics for Twitter Users

Presenter: Quan Yuan (SCE)
Supervisor: Asst. Prof. Gao Cong
Prof. Nadia Magnenat- Thalmann

Nanyang Technological University
KDD 2013
Motivation

- Micro-blogging services (e.g., Twitter) and location-based social networks (LBSNs, e.g., Foursquare) have generated a great number of geotagged short text messages.

- A short message contains a user id, a text message, posting time and a venue.

- Such short messages offers a good opportunity to study the behaviors of individuals (who) from geographical location (where), time (when) and activity (what).
Related Work

- Previous studies focused on at most three factors of Who, Where, When, and What

  - **Where What**: geographical topic modeling
    - *Yin et al.* Geographical topic discovery and comparison. *In WWW 2011*
    - *Chinatown*: dinning, *Raffles*: sightseeing, etc..

  - **Where When What**: geographical event detection
    - *Sakaki et al.* Earthquake shakes twitter users: real-time event detection by social sensors. *In WWW 2010*
    - *China* @ Jan 28st: lunar new year

  - **Who Where When**: modeling spatio-temporal mobility behaviors of individual users
    - *Cho et al.* Friendship and mobility: user movement in location-based social networks. *In KDD 2011*
    - *Tony*: office@2:00 pm, home@9:00pm

  - **Who Where What**: geographical topic profiling of users
    - *Hong et al.* Discovering geographical topics in the twitter stream. *In WWW 2012*
    - *Tony @ Jurong Point*: shopping, dinning
Overview: Region

- A tweet $d$ is modeled as a five-tuple $\{u_d, l_d, w_d, t_d, s_d\}$
  - $u$: user, $l=$\{id, coordinate\}: venue id and geographical coordinate
  - $w$: words, $t=$\{hh:mm:ss\}: time in a day, $s$: workday/weekend

- Intuitions:
  - An individual $u$’s mobility centres at different personal geographical regions $r$ (e.g., home region, work region, shopping region, etc).
  - The region where a user stays is influenced by time (time in a day $t$ & day of a week $s$).
  - Eg: Tony: 2:00 pm, weekday – NTU; 9:00 pm, weekend – Home
  - Region: Gaussian distribution over latitude/longitude.
Overview: Topic

- A tweet is modeled as a five-tuple \( d = \{u_d, l_d, w_d, t_d, s_d\} \)
  - \( u \): user, \( l = \{id, coordinate\} \): venue id and geographical coordinate
  - \( w \): words, \( t = \{hh:mm:ss\} \): time in a day, \( s \): workday/weekend

- Intuitions:
  - Different users \( u \) and regions \( r \) have different preferences over topics \( z \), \( \{P(z|u)\}, \{P(z|r)\} \)
  - The topics \( z \) of a user \( u \) at a place are influenced by region \( r \) and \( u \)'s topic preference.
  - Eg.: Home: cooking, music. Tony: cooking, reading
    Tony @ Home: cooking

Graphical Model
A tweet is modeled as a five-tuple \( d = \{u_d, l_d, w_d, t_d, s_d\} \)

- **u**: user
- **l**: venue id and geographical coordinate
- **w**: words
- **t**: time in a day
- **s**: workday/weekend

**Intuitions:**

- When choosing a venue \( l \) to visit, the user considers the topic requirement and the region.
  \[ P(l|r,z) = \alpha \cdot P(id|z) + (1-\alpha) \cdot P(coor|r) \]
- E.g.: Tony, shopping: fairprice
- The selection of words are also influenced by topic and region.
  \[ P(w|r,z) = \beta \cdot P(w|z) + (1-\beta) \cdot P(w|r) \]
Parameter Estimation

- The likelihood:
  \[
  \sum_d \sum_z \sum_r \log \left( \sum p(u_d, r, z, s_d, t_d, \ell_d, w_d) \right) = \sum_d \sum_z \sum_r p(u_d)p(s_d|u_d)p(t_d|u_d, s_d)p(r|u_d, s_d, t_d) \\
  \prod_{w \in w_d} (\lambda p(w|z) + (1 - \lambda)p(w|r)) c(w, w_d) [\kappa p(\ell_d|z) + (1 - \kappa)p(\ell_d|r)] p(w_d|r, z)
  \]

- We use Expectation-Maximization (EM) to estimate parameters

  - E-step: 
    \[
    p(r, z|d) = \frac{p(d, r, z)}{p(d)} = \frac{p(d, r, z)}{\sum_r \sum_z p(d, r, z)}
    \]

  - M-step:
    \[
    p(r|u, s) = \frac{\sum_{d \in \mathcal{D}_{u,s}} \sum_z p(r, z|d)}{\sum_{d \in \mathcal{D}_{u}} \sum_z \sum_{r'} p(r', z|d)}
    \]
    \[
    p(z|u, r) = \frac{\sum_{d \in \mathcal{D}_{u}} p(r, z|d)}{\sum_{d \in \mathcal{D}_{u}} \sum_z p(r, z'|d)}
    \]
    \[
    \mu_{u,s,r} = \frac{\sum_{d \in \mathcal{D}_{u,s}} \sum_z p(r, z|d) \cdot cd_{\ell_d}}{\sum_{d \in \mathcal{D}_{u,s}} \sum_z p(r, z|d)}
    \]
    \[
    \Sigma_{u,s,r} = \frac{\sum_{d \in \mathcal{D}_{u,s}} \sum_z p(r, z|d) \cdot (cd_{\ell_d} - \mu_{u,s,r})^T (cd_{\ell_d} - \mu_{u,s,r})}{\sum_{d \in \mathcal{D}_{u,s}} \sum_z p(r, z|d)}
    \]
    \[
    p(\ell|z) = \frac{\sum_{d \in \mathcal{D}_\ell} \sum_{r'} p(r, z|d)}{\sum_{d \in \mathcal{D}_\ell} \sum_{r'} \sum_r p(r, z'|d)}
    \]
Datasets and Evaluation Tasks

- **Datasets:**
  - **WW:** 89,007 world-wide tweets, 3,883 users, 60,962 venues
  - **USA:** 171,768 microblogs in USA, 4,122 users, 35,989 venues

- **Venue prediction for tweets**
  - Given text contents, user ids and posting time, predict the most likely venue at which the tweet is posted.
  - Rank candidate venues $l$ by $p(l \mid u, s, t)$

- **User prediction:**
  - Predict the likelihood of a user visiting a venue at a given time.
  - Rank candidate users $u$ by $p(u \mid l, s, t)$

- **Venue prediction for user:**
  - Predict the place where a user stays at a given time.
  - Rank candidate venues $l$ by $p(l \mid u, s, t, w)$
Venue prediction for tweets

Metric:
- Prediction accuracy (Acc): percentage of tweets whose predicted venues are the true venues.
- Average error distance (Dis): the average geographical distance between the predicted and true venues for all tweets.
- Larger Acc and smaller Dis indicate better performance.

Baselines:
- KL-divergence based method (KL) W. Li et al. CIKM 2011
- Topic+Region (TR) L. Hong et al. WWW 2012
- Who+Where+What (W3)
- Who+Where+When+What (W4)

<table>
<thead>
<tr>
<th>Factors in modeling</th>
<th>KL</th>
<th>TR</th>
<th>W³</th>
<th>W⁴</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who (User)</td>
<td>×</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Where (Geo)</td>
<td>×</td>
<td>G1bR</td>
<td>PsnR</td>
<td>PsnR</td>
</tr>
<tr>
<td>When (Time)</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>√</td>
</tr>
<tr>
<td>What (Words)</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
</tbody>
</table>
Venue prediction for tweets

- Prediction results:
  - W4 outperforms KL and TR by more than 80% in terms of both metrics.
  - W3 utilizes the same information as does TR, but gains better results
    - Regions are personal in W3, but global in TR
  - W4 achieves best results
    - Time factor is important
User Prediction and Venue Prediction

- Metric: Accuracy (Acc)
- User Prediction:
  - Predict the likelihood of a user visiting a venue at a given time

<table>
<thead>
<tr>
<th>Acc</th>
<th>WW</th>
<th>USA</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMM</td>
<td>0.4163</td>
<td>0.4021</td>
</tr>
<tr>
<td>W4</td>
<td>0.5063</td>
<td>0.5863</td>
</tr>
</tbody>
</table>

- Venue Prediction for User
  - Predict the place where a user stays at a given time.

<table>
<thead>
<tr>
<th>Acc</th>
<th>WW</th>
<th>USA</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMM</td>
<td>0.0423</td>
<td>0.1102</td>
</tr>
<tr>
<td>W4</td>
<td>0.0776</td>
<td>0.2953</td>
</tr>
</tbody>
</table>

- W4 outperforms PMM by more than 20% and 45% on the two datasets, respectively
Results of Other Tasks

- **Activity prediction:**
  - Predict word of a user at a given time

<table>
<thead>
<tr>
<th>time</th>
<th>words</th>
</tr>
</thead>
<tbody>
<tr>
<td>14:30 weekday</td>
<td>break work coffee resting gym international</td>
</tr>
<tr>
<td>10:00 weekend</td>
<td>good morning home breakfast shopping eat</td>
</tr>
</tbody>
</table>

- **User mobility**

- **Representative words for topics**

<table>
<thead>
<tr>
<th>Topic</th>
<th>Representative words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Home</td>
<td>family fun offroad rental home love</td>
</tr>
<tr>
<td>Dining</td>
<td>lunch dinner birthday breakfast drinks eat</td>
</tr>
<tr>
<td>Nightlife</td>
<td>night happy singing playing dance football</td>
</tr>
<tr>
<td>Work</td>
<td>working tonight coffee tired money Friday</td>
</tr>
<tr>
<td>Holiday</td>
<td>Christmas friends holiday merry celebrating choir</td>
</tr>
</tbody>
</table>
Conclusion

- The large availability of geo-tagged tweets enables us to study individuals’ mobility behaviors user, geolocation, time, and activity factors.

- We propose W4 (**Who Where When What**) to model the interactions of all the four factors in a unified way to better understand individuals’ behaviors.

- Experimental results on two real-world datasets show that the proposed method outperforms baselines on various applications.
Q & A?

Thank You!
Publications


