Spatiotemporal Event Forecasting in Social Media

Liang Zhao, Feng Chen, Chang-Tien Lu, Naren Ramakrishnan.
SIAM Data Mining 2015, Vancouver, BC, Apr-May (to appear)
Spatiotemporal Events

Influenza outbreak on Week 47 ending Nov 22, 2014 in southern region

Civil unrest events on Mar 17, 2013 in Brazil
Social Media as Surrogate of Spatiotemporal Events

Mexico presidential election on July 1, 2012

Why? And what preceded this protest?

Protests on July 7, 2012

July 1

July 7
Social Media for Event Forecasting

This is just beginning ... this has not been a democratic election

Today #tauro do not hesitate to raise your voice in protest against injustice. Who comes to march?

Let us go to the Mega March on July 7 to Zócalo Angel on 3:00 pm. spread the word

Complaining  | Planning  | Dissemination  | Protest event

- tweet volume>500
- 50<tweet volume<500
- 0<tweet volume<50
- Location of event
Challenge 1: Capturing evolutionary structural contexts

- **Background words**: This is just beginning ... this has not been a democratic election
- **Event specific words**: Today #tauro do not hesitate to raise your voice in protest against injustice. Who comes to march?
- **Topic 1 words**: Let us go to the MegaMarch on July 7 to Zócalo Angel on 3:00 pm. spread the word

Timeline:
- **Complaining**: July 1, 2012
- **Planning**: July 3, 2012
- **Dissemination**: July 5, 2012
- **Protest event**: July 7, 2012
Challenge 2: Modeling mixed type observations

- Complaining: July 1, 2012
  - This is just beginning... this has not been a democratic election

- Planning: July 3, 2012
  - Today #tauro do not hesitate to raise your voice in protest against injustice. Who comes to march?

- Dissemination: July 5, 2012
  - Let us go to the Mega March on July 7 to Zócalo Angel on 3:00 pm. spread the word

- Protest event: July 7, 2012

Legend:
- tweet volume>500
- 50<tweet volume<50
- 0<tweet volume<50
- Location of event
Challenge 3: Utilizing prior geographical knowledge

Crime events happen more often in regions with bad security

Earthquakes occur more frequently on continent plate border

Civil unrest events occur more frequently on large cities
Our contribution

• Generative model of the event development progression from social media data.
• Effective algorithm for model parameter inference.
• Event forecasting algorithm by sequence likelihood calculation.
• Multiple types of events data to demonstrate the effectiveness of forecasting.
Our overall model

Spatial burstiness model

Structural context model

Event temporal progression
Spatiotemporal burstiness modeling

- 1. geographical prior
- 2. bi-variate Gaussian to model the correlation of in-out count.

$$ r := (r_{in}, r_{out}) $$
Categorical context in tweets

Background words

- This is just beginning ... this has not been a democratic election

Event specific words

- Today #tauro do not hesitate to raise your voice in protest against injustice. Who comes to march?

Topic 1 words

- Let us go to the MegaMarch on July 7 to Zócalo Angel on 3:00 pm. spread the word

Topic 2 words

Timeline:
- Complaining: July 1, 2012
- Planning: July 3, 2012
- Dissemination: July 5, 2012
- Protest event: July 7, 2012
Context semantics modeling

• 1. Event-specific words
• 2. Topic model
Context semantics modeling

- 1. Event-specific words
- 2. Topic model
Inference

- Objective

\[
p(W, X, Y, Z, \mu, \Sigma, r^{in}, r^{out}|\pi, A, \Psi, \Phi, \theta, \Theta_0)
= \prod_s p(Z_{s,1}|\pi) \cdot \prod_s \prod_{t=2}^{T} p(Z_{s,t}|Z_{s,t-1}, A)
\cdot \prod_s \prod_{t=1}^{T} \prod_{n=1}^{N} p(W_{s,t,n}, Y_{s,t,n}, X_{s,t,n}|Z_{s,t}, \Psi, \Phi, \theta)
\cdot \prod_s \prod_{t=1}^{T} \prod_{i=1}^{N} p(r_{s,t}^{in}, r_{s,t}^{out}|\mu_l, \Sigma_l, Z_{s,t}) p(\mu_l, \Sigma_l|\Theta_0)
\]

- E step: fix \( \mu_l, \Sigma_l, \Psi, \Phi, \theta \), optimize \( \mathbb{E}[p(Z_{s,t} = k)] \) and \( \mathbb{E}[p(Z_{s,t} = j|Z_{s,t} = k)] \)

- M step: fix \( \mathbb{E}[p(Z_{s,t} = k)] \) and \( \mathbb{E}[p(Z_{s,t} = j|Z_{s,t} = k)] \), optimize \( \mu_l, \Sigma_l, \Psi, \Phi, \theta \).
Event Forecasting: Sequence Classification

• Using the proposed generative process, two models $C_1$ and $C_2$ are built.

• $C_1$ characterizes the development progression leading to an event, $C_2$ characterizes the one leading to no event.

• Denote new sequence as $s_t$, if $p(C_1|s,l) > \varepsilon \cdot p(C_2|s,l)$, there is event, otherwise, there is not.

\[
p(C_1|s,l) = p(s|C_i) \cdot p(C_i|l)/p(s)
\]

\[
p(s|C_i) = \max_{\{Z_t\}_1^T, \theta^R, n^R, n^B} \ln p(s, Z_1, \ldots, Z_T|C_i)
\]
Sequence likelihood calculation: modified Viterbi

Sequence likelihood: \( p(s|C_i) = \max_{\{Z_t\}_t^T, \theta^R, n_R, n_B} \ln p(s, Z_1, \ldots, Z_T|C_i) \)

\[
\omega_t = \max_{\theta^R_{s,t}, n_R, n_B} \ln p(s_t|Z_t, C_i) + \max_{Z_{t-1}} \{ \ln p(Z_t|Z_{t-1}) + \omega_{t-1} \}
\]

\[
\theta^R_{s,t,i} = \frac{n^R_{s,t,i}}{\xi \cdot \Psi^R_{k,1}} \quad \sum_{i} n^R_{s,t,i} \cdot \log \theta^R_{s,t,i} + \sum_{j} n^B_{j} \cdot \log \theta^B_{j} \\
\text{s.t.} \sum_{i} \theta^R_{s,t,i} = 1, n^R_{s,t,w} + \sum_{j} n^B_{j} = \xi_w, n^R_{s,t,w} \geq 0, n^B_{j} \geq 0, \sum_{i} n^B_{j} = \xi \cdot \Psi^R_{k,1}
\]
Experiments: Datasets

• Civil unrest:
  • Time duration: 2013-01-01 - 2013-06-01
  • #raw tweets: 32,459,668
  • #processed tweets: 57,856
  • #events: 726

• Flu outbreaks:
  • Time duration: 2011-01-01 - 2013-12-31
  • #raw tweets: 8,627,664,399
  • #processed tweets: 2,252,436
  • #events: 102
Performance comparisons

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Metric</th>
<th>Baseline</th>
<th>ARX</th>
<th>LR</th>
<th>LDA-LR</th>
<th>KDE-LDA-LR</th>
<th>Proposed algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>civil unrest data</td>
<td>precision</td>
<td>0.44</td>
<td>0.26</td>
<td>0.7</td>
<td>0.31</td>
<td>0.42</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>recall</td>
<td>0.59</td>
<td>0.43</td>
<td>0.18</td>
<td>0.7</td>
<td>0.69</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>f1-score</td>
<td>0.5</td>
<td>0.32</td>
<td>0.29</td>
<td>0.43</td>
<td>0.52</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>runtime per day (sec)</td>
<td>$10^{-3}$</td>
<td>$10^{-3}$</td>
<td>0.001</td>
<td>0.005</td>
<td>0.005</td>
<td>0.32</td>
</tr>
<tr>
<td>flu data</td>
<td>precision</td>
<td>0.28</td>
<td>0.14</td>
<td>0.64</td>
<td>0.27</td>
<td>0.78</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>recall</td>
<td>0.39</td>
<td>0.66</td>
<td>0.31</td>
<td>0.55</td>
<td>0.32</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>f1-score</td>
<td>0.33</td>
<td>0.23</td>
<td>0.41</td>
<td>0.36</td>
<td>0.46</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>runtime per day (sec)</td>
<td>$10^{-3}$</td>
<td>0.001</td>
<td>0.01</td>
<td>0.02</td>
<td>0.03</td>
<td>2.1</td>
</tr>
</tbody>
</table>
Sensitivity analysis: #states

![Graph showing F1 score vs number of latent states for Flu dataset and Civil unrest dataset. The graph indicates a peak in F1 score around 7 latent states for both datasets.]
Sensitivity analysis: \#topics
Sensitivity analysis: cost ratio

(a) Civil unrest dataset

(b) Flu dataset
Thank you