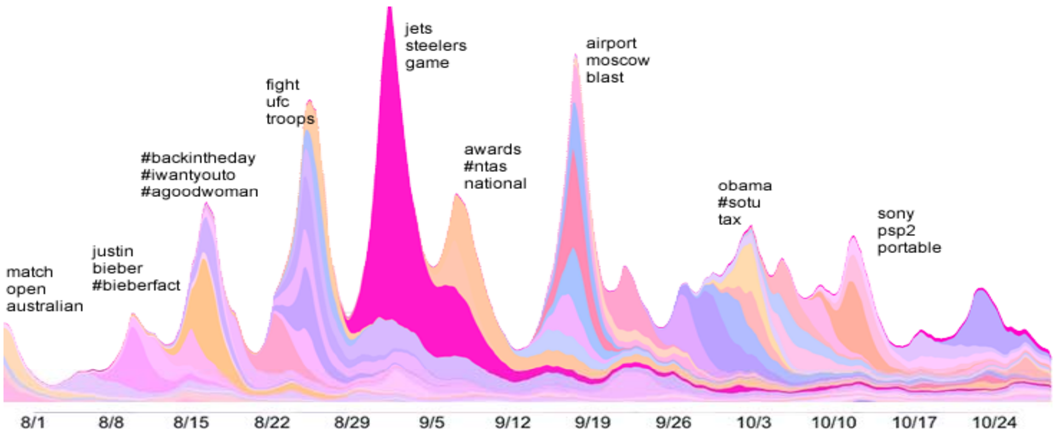
A Probabilistic Model for Bursty Topic Discovery in Microblogs

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Jun Xu, Xueqi Cheng
CAS Key Laboratory of Web Data Science and Technology

Bursty Topics in Microblogs



Bursty topics: novel topics attracting wide interest (hot events, activities, discussions)

Valuable information

public opinion analysis

business intelligence

news clues tracking

Message recommendation

…

Problems & Challenges

- Microblog Posts are very short
  - Conventional topic models (e.g., LDA and PLSA) are not effective over short texts
  - How to discover topics in such short texts?

- Microblog Posts are very diverse and noisy
  - Lots of pointless babbles, daily chatting and other non-bursty content
  - How to distinguish bursty topics from other topics?

Our Work

- We propose a probabilistic model to solve the two challenges in a principled and effective way

How to learn topics over short texts?

Exploit the rich global word co-occurrence to learn topics (following our previous work biterm topic model)

How to distinguish bursty topics from non-bursty content?

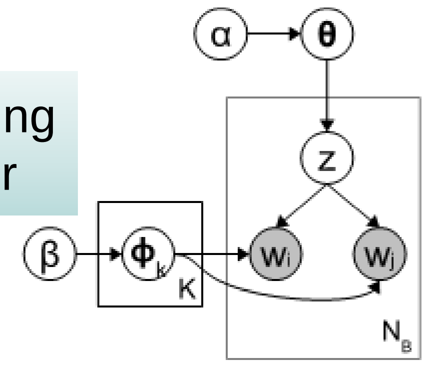
Exploit the burstiness of biterms as prior knowledge for bursty topic discovery

Biterm Topic Model (BTM) for Short Texts

(Yan et.al WWW'13)

- LDA will encounter the data sparsity problem when docs are short
- BTM models the generation of biterm
  - Drawn a topic z from a global topic distribution θ
  - draw two word from the topic z

Co-occurring word pair



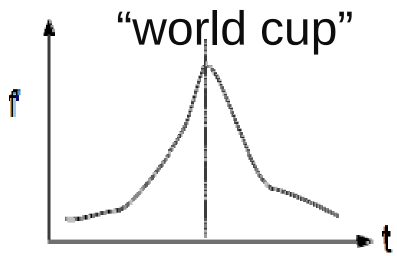
BTM

- BTM can better learn topics over short texts
  - Directly model the word co-occurrence
  - fully exploit the rich global word co-occurrence to overcome the data sparsity problem in short documents

But BTM learns general topics rather than bursty topics ☹

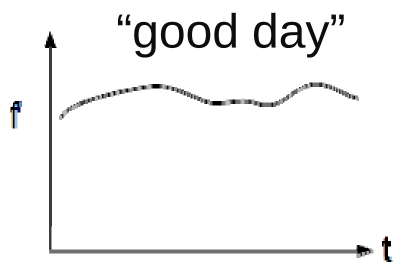
Observations

- A **bursty biterm** is **more** likely to be generated from some bursty topic



bursty topic about
the World Cup 2014

- A **non-bursty biterm** is **less** likely to be generated from any bursty topics



Non-bursty topic

Bursty Probability of a Biterm

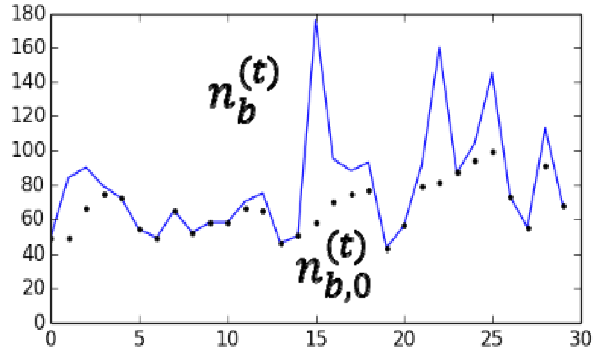
- How likely a biterm will be generated from some bursty topic?

- Suppose each biterm is generated either from some bursty topic or other non-bursty topic

Count in non-bursty topics

Count in bursty topics

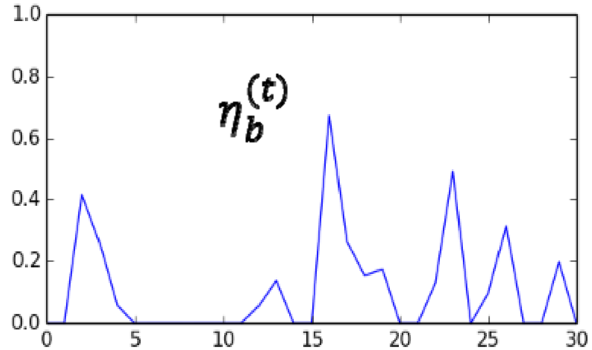
Actual count $n_b^{(t)} = n_{b,0}^{(t)} + n_{b,1}^{(t)}$



  - $n_{b,0}^{(t)}$ can be estimated by the average count in the last S time slices

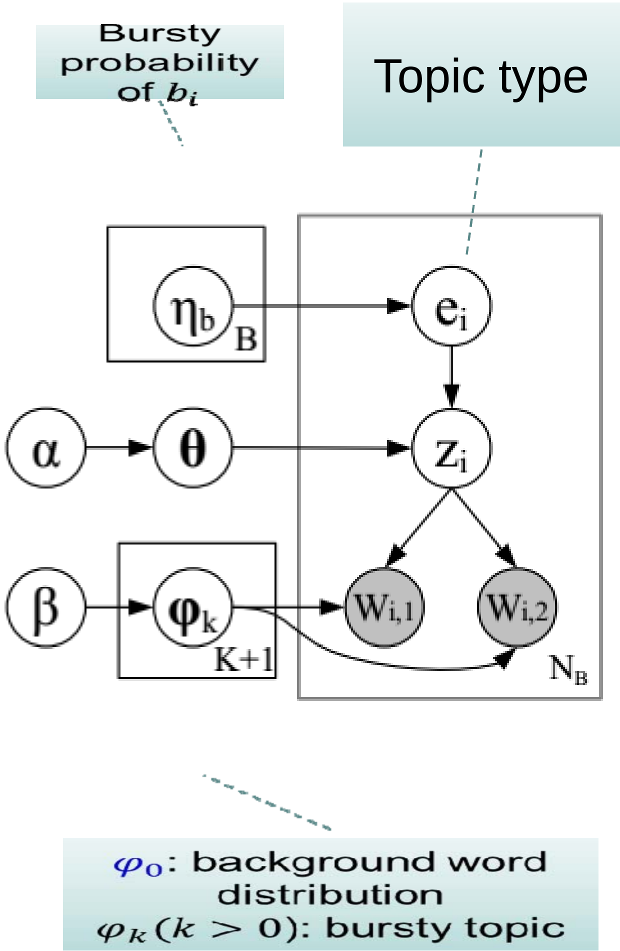
$$n_{b,0}^{(t)} = \min\left(\frac{1}{S}\sum_{s=1}^{S} n_b^{(t-s)}, n_b^{(t)}\right)$$

$$\eta_b^{(t)} = \frac{n_{b,1}^{(t)}}{n_b^{(t)}} = 1 - \frac{n_{b,0}^{(t)}}{n_b^{(t)}}$$



Bursty Biterm Topic Model (BBTM)

1. For the collection,
   - draw a bursty topic distribution $\boldsymbol{\theta} \sim \mathrm{Dir}(\alpha)$
   - draw a background word distribution $\boldsymbol{\phi}_0 \sim \mathrm{Dir}(\beta)$
2. For each bursty topic $k \in [1, K]$,
   - draw a word distribution $\boldsymbol{\phi}_k \sim \mathrm{Dir}(\beta)$
3. For each biterm $b_i \in \mathbb{B}$
   - draw $e_i \sim \mathrm{Bern}(\eta_{b_i})$
   - If $e_i = 0$,
     - draw two words $w_{i,1}, w_{i,2} \sim \mathrm{Multi}(\boldsymbol{\phi}_0)$
   - If $e_i = 1$,
     - draw a bursty topic $z \sim \mathrm{Multi}(\boldsymbol{\theta})$
     - draw two words $w_{i,1}, w_{i,2} \sim \mathrm{Multi}(\boldsymbol{\phi}_z)$



Parameter Inference by Gibbs Sampling

- Randomly assign a topic for each biterm
- Repeatedly update the topic for each biterm in a sequential way until convergence

$$P(e_i = 1, z_i = k | \mathbf{e}^{\neg i}, \mathbf{z}^{\neg i}, \mathbb{B}, \alpha, \beta, \boldsymbol{\eta}) \propto \eta_{b_i} \cdot \frac{(n_k^{\neg i} + \alpha)}{(n_\cdot^{\neg i} + K\alpha)} \cdot \frac{(n_{k,w_{i,1}}^{\neg i} + \beta)(n_{k,w_{i,2}}^{\neg i} + \beta)}{(n_{k,\cdot}^{\neg i} + W\beta)(n_{k,\cdot}^{\neg i} + 1 + W\beta)}$$

Probability of a biterm b_i assigned to the kth bursty topic

Bursty probability of b_i

Popularity of the topic

Closeness between the two words and the bursty topic

BBTM always choose the bursty and relevant biterms to construct bursty topics

Experiments

- Trec Tweets2011 Collection
  - 17 days：2011.1.23-2011.2.8
  - 4,230,578 tweets, 98,857 distinct terms

- Baselines
  - Twevent: state-of-the-art heuristic-based approach
  - oLDA : online LDA + post-processing
  - UTM: User Temporal Topic Model, state-of-the-art model-base approach
  - IBTM：BTM + post-processing
  - BBTM-S: draw e_i in the first iteration and then fix it

Accuracy of Bursty Topics

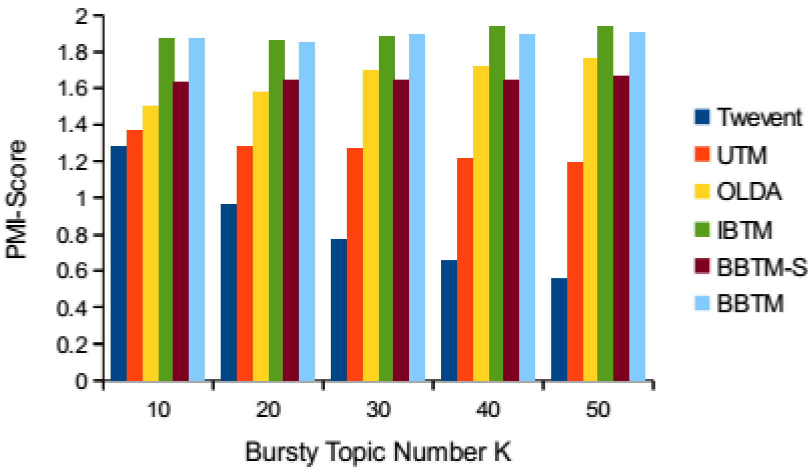
- Metric : $Precision@K_1$

Method	P@10	P@30	P@50
Twevent	0.592	0.681	0.636
UTM	0.565	0.488	0.453
OLDA	0.231	0.217	0.185
IBTM	0.300	0.325	0.297
BBTM-S	0.785	0.832	0.790
BBTM	**0.810**	**0.865**	**0.842**

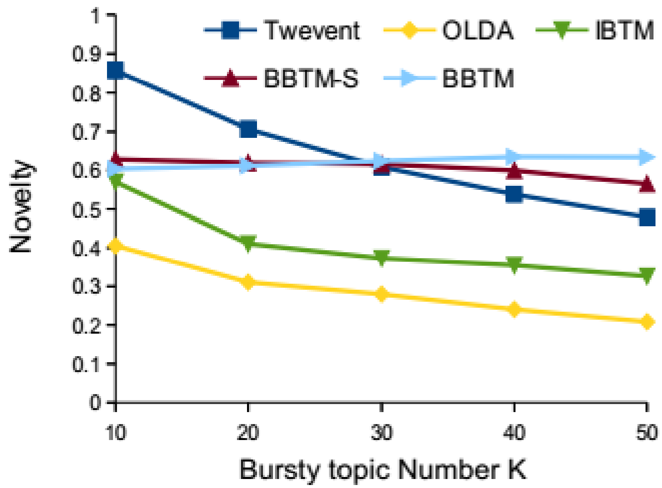
k	The 10 most probable words	$\hat{\theta}_k$
2	police officers shot shooting detroit twitter adam suspect year revenue (Two St. Petersburg police officers were shot and killed)	0.036
11	airport moscow police news killed people dead blast suicide explosion (Deadly suicide bombing hits Moscow's Domodedovo airport)	0.057
15	open #ausopen nadal australian murray mike tomlin cloud #cloud avril (Australian Open Tennis Championships 2011)	0.015
25	jack lalanne fitness 96 dies guru rip died age dead (Jack LaLanne: US fitness guru who last ate dessert in 1929 dies aged 96)	0.044
26	court emanuel rahm chicago ballot mayor mayoral run appellate rules (Court tosses Emanuel off Chicago mayoral ballot)	0.024

Coherence and Novelty of Bursty Topics

- Coherence
  - More relevant the top words in topics, the coherence is higher



- Novelty
  - More words are overlap between two time slices, the novelty is lower



Efficiency

K	UTM	OLDA	IBTM	BBTM-S	BBTM
10	4.24	1.84	4.66	**0.03**	1.57
20	6.02	2.61	5.97	**0.06**	2.89
30	7.84	3.28	7.24	**0.09**	4.40
40	9.79	4.02	8.54	**0.13**	5.71
50	11.63	4.83	9.99	**0.17**	7.24

Table 4: Time cost (second) per iteration.

- BBTM-S costs much less time than other methods
  - since it only used a subset of biterms for training
- BBTM is more efficient than IBTM and UTM.
  - since they waste time to learn non-bursty topics

Summary

- We propose the bursty biterm topic model for bursty topic discovery in microblogs
  - It exploits the rich global word co-occurrence for better topic learning over short texts
  - It exploits the burstiness of biterms to distill bursty topics automatically

- Furture works
  - Improve the estimation of bursty probability
  - Improve topic representations with external data

Thank You！

Code: https://github.com/xiaohuiyan/BurstyBTM
Our related work: http://shortext.org/