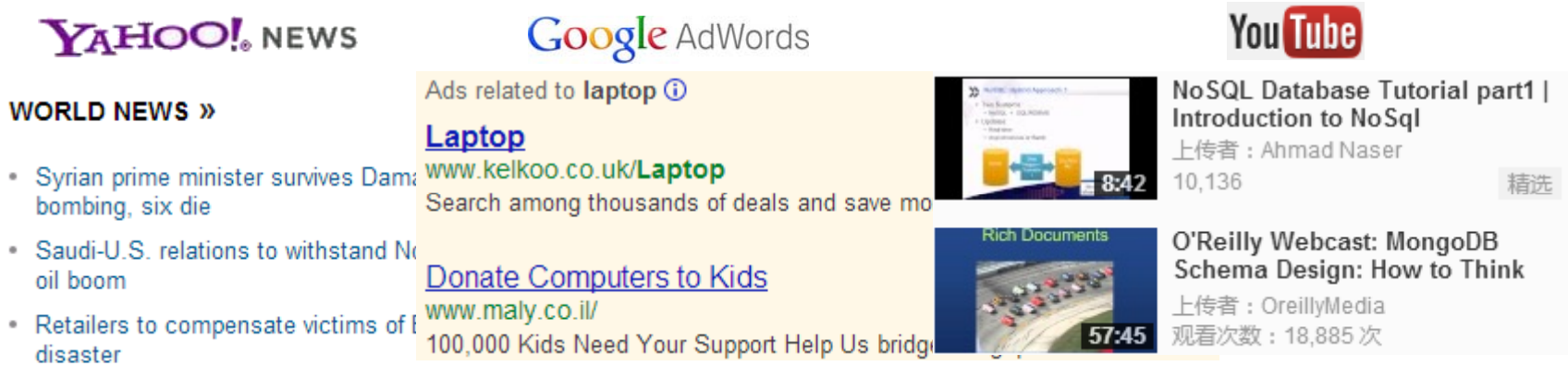
A Biterm Topic Model for Short Texts

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Xueqi Cheng
Institute of Computing Technology,
Chinese Academy of Sciences

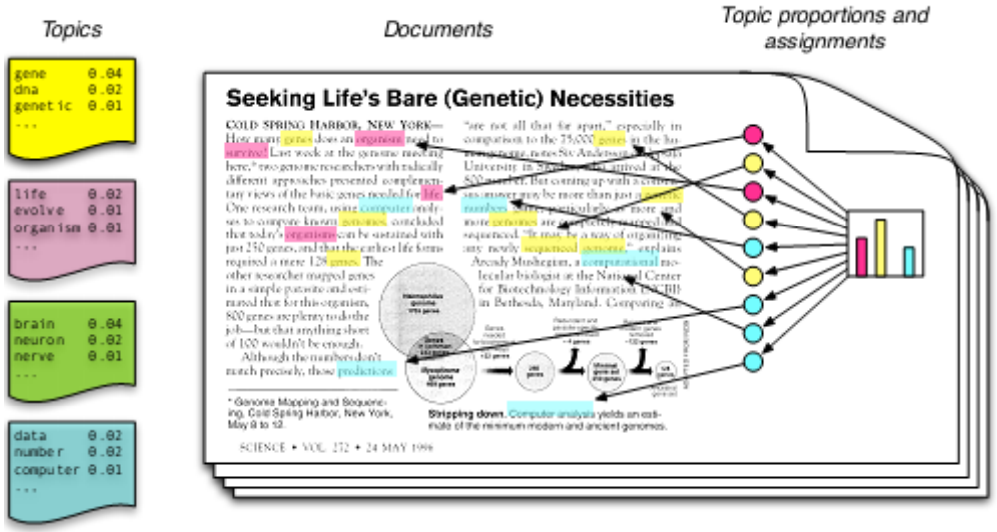
Short Texts Are Prevalent on Today's Web



Background

- Understanding the topics of short texts is important for many tasks
  - content recomendation
  - user interest profiling
  - content characterizing
  - emerging topic detecting
  - semantic analysis
  - ...
- This work originates from a browsing recomendation project

Topic Models



From Blei

- Model the generation of documents with latent topic structure
  - a topic ~ a distribution over words
  - a document ~ a mixture of topics
  - a word ~ a sample drawn from one topic
- Previous studies mainly focus on normal texts

Problem on Short Texts: Data Sparsity

- Word counts are not discriminative
  - normal doc: topical words occur frequently
  - short msg: most words only occur once
- Not enougth contexts to identify the senses of ambiguous words
  - normal doc: rich context, many relevant words
  - short msg: limited context, few relevant words
- The severe data sparsity makes conventional topic models less effective on short texts

Previous Approaches on Short Texts

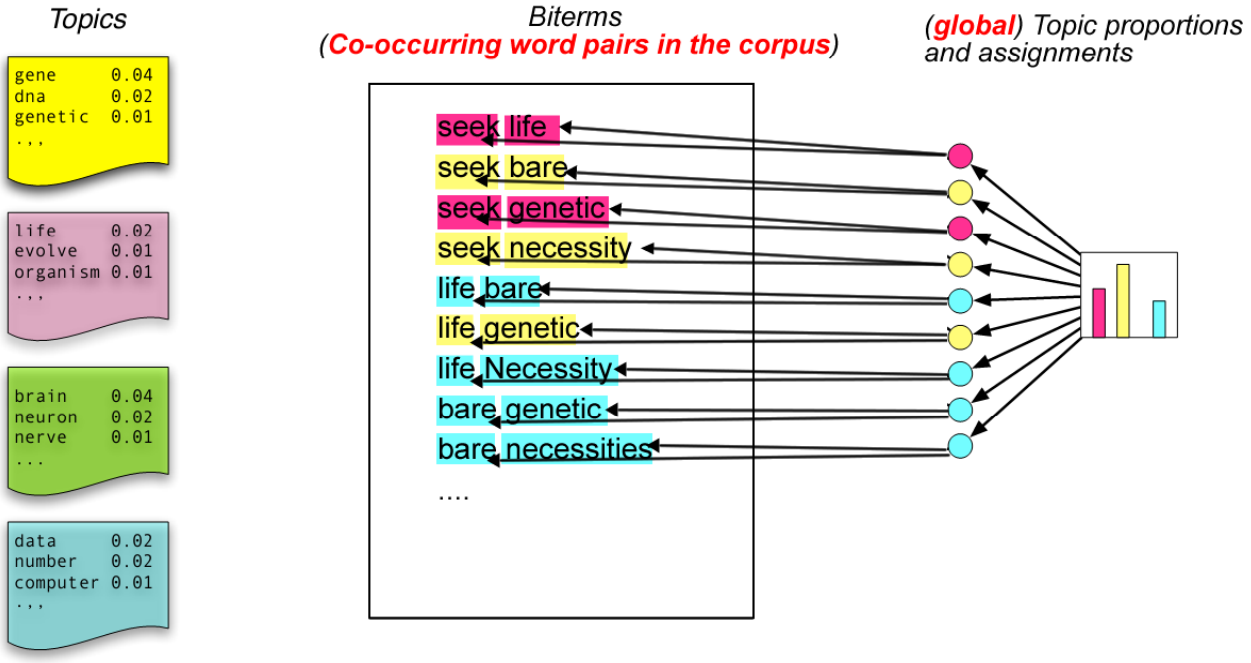
- Document aggregation
  - e.g. aggregating the tweets published by the same users
  - heuristic, not general
- Mixture of unigrams
  - each document has only one topic
  - too strict assumption, peaked posteriors P(z|d)
- Sparse topic models
  - add sparse constraints on the distribution over topics in a document, e.g. Focused Topic Model
  - too complex, easy to overfit

Key Idea

- A Topic is basically a group of correlated words and the correlation is revealed by word co-occurrence patterns in documents
  - **why not directly model the word co-occurrences for topic learning?**

- Conventional Topic models suffer from the problem of severe sparse patterns in short documents
  - **why not use the rich global word co-occurrence patterns for better revealing topics instead?**

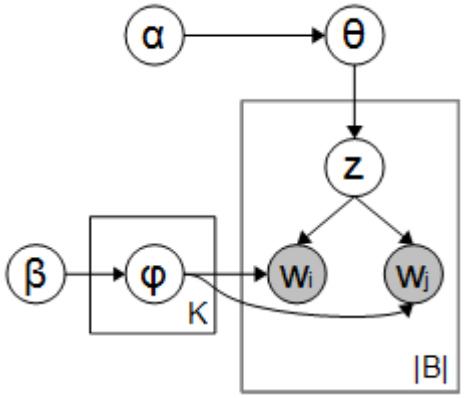
Biterm Topic Model (BTM)



- Model the generation of biterms with latent topic structure
  - a topic ~ a distribution over words
  - a corpus ~ a mixture of topics
  - a biterm ~ two words drawn from one topic

Generation Procedure of Biterms

1. For each topic z
   (a) draw a topic-specific word distribution $\boldsymbol{\phi}_z \sim Dir(\boldsymbol{\beta})$
2. Draw a topic proportion vector $\boldsymbol{\theta} \sim Dir(\boldsymbol{\alpha})$ for the whole collection
3. For each biterm $\mathbf{b}$
   (a) draw a topic assignment $z \sim Multi(\boldsymbol{\theta})$
   (b) draw two words: $w_i, w_j \sim Mulit(\boldsymbol{\phi}_z)$



Inferring Topics in a Document

- Assumption
  - the topic proportions of a document equals to the expectation of the topic proportions of biterms in it

$$P(z|d) = \sum_b P(z|b)P(b|d)$$

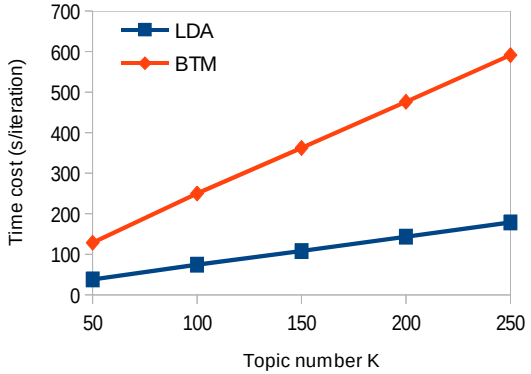
where

$$P(z|b) = \frac{P(z)P(w_i|z)P(w_j|z)}{\sum_z P(z)P(w_i|z)P(w_j|z)}, \qquad P(b|d) = \frac{n_d(b)}{\sum_b n_d(b)}$$

Parameters Inference

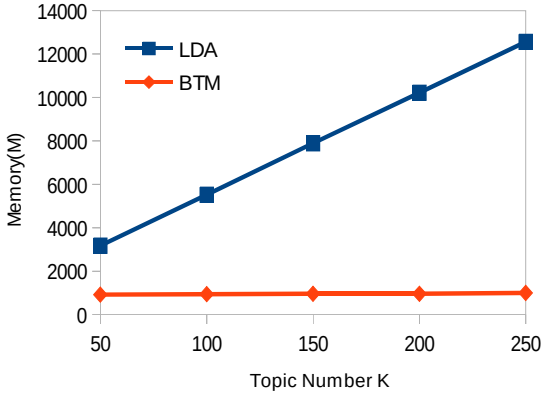
- Gibbs Sampling
  - sample topic for each biterm

$$P(z|\mathbf{z}_{-b}, B, \alpha, \beta) \propto (n_z + \alpha)\frac{(n_{w_i|z} + \beta)(n_{w_j|z} + \beta)}{(\sum_w n_{w|z} + M\beta)^2}$$



  - parameters estimate

$$\phi_{w|z} = \frac{n_{w|z} + \beta}{\sum_w n_{w|z} + M\beta},$$
$$\theta_z = \frac{n_z + \alpha}{|B| + K\alpha},$$



- BTM is more memory-efficient than LDA

Experiments: Datasets

	Tweets2011 (short text)	Question (short text)	20Newsgroup (normal text)
#documents	4,230,578	189,080	18,828
#words	98,857	26,565	42,697
#users	2,039,877	-	-
#categories	-	35	20
avg doc length (after pre-processing)	**5.21**	**3.94**	**97.20**

Experiments: Tweets2011 Collection

- Topic quality
  - Metric: average coherence score (Mimno'11) on the top T words
    - A larger coherence score means the topics are more coherent

T	5	10	20
LDA	-55.0 ± 0.4	-236.4 ± 2.0	-1015.7 ± 5.9
LDA-U	-54.2 ± 0.8	-234.8 ± 1.1	-1009.4 ± 4.4
Mix	-53.8 ± 0.1	-233.0 ± 1.4	-1007.6 ± 6.7
BTM	$\mathbf{-52.4 \pm 0.1}$	$\mathbf{-227.8 \pm 0.3}$	$\mathbf{-990.2 \pm 3.8}$

D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic oherence in topic models. EMNLP 2011

Experiments: Tweets2011 Collection

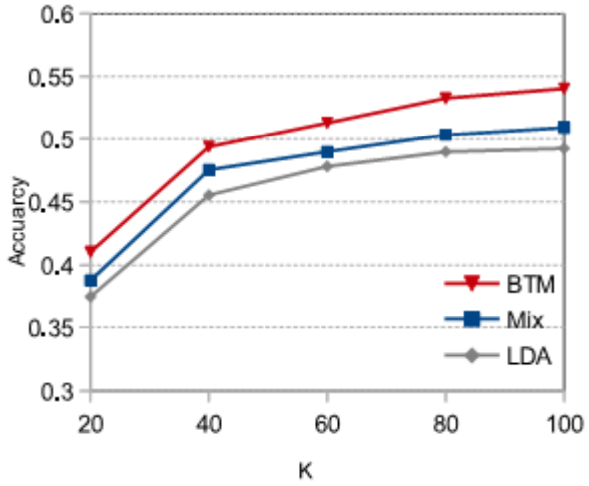
- Quality of topic proportions of documents (i.e. P(z|d))
  - select 50 frequent and meanful hashtags as class labels
  - organize documents with the same hashtag into a cluster
  - measure: H score
    - smaller value indicates better agreement with human labeled classes

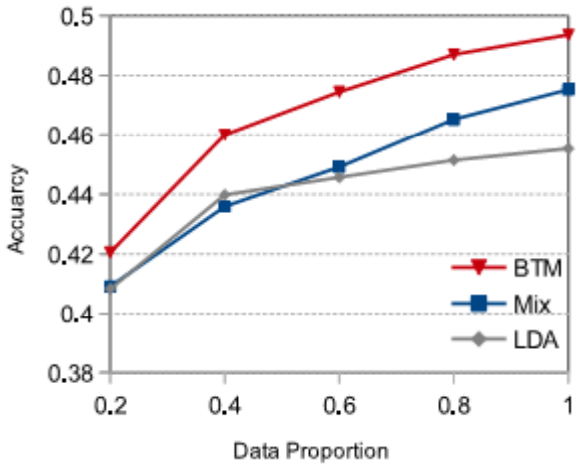
$$H = \frac{IntraDis(C)}{InterDis(C)}$$

Method	H score	Significant differences
LDA	0.576 ± 0.007	
LDA-U	0.564 ± 0.011	>LDA*
Mix	0.503 ± 0.008	>LDA-U**>LDA***
BTM	0.474 ± 0.005	>Mix***>LDA-U***>LDA***

Experiments: Question Collection

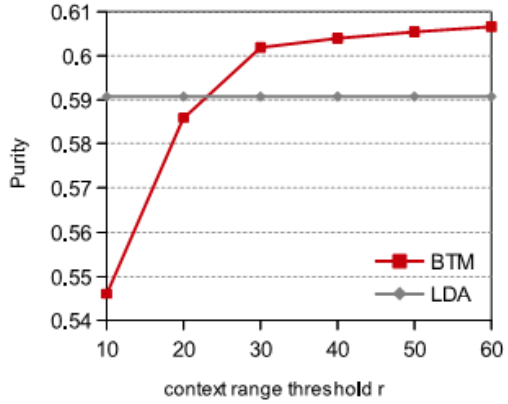
- Evaluated by document classification (linear SVM)

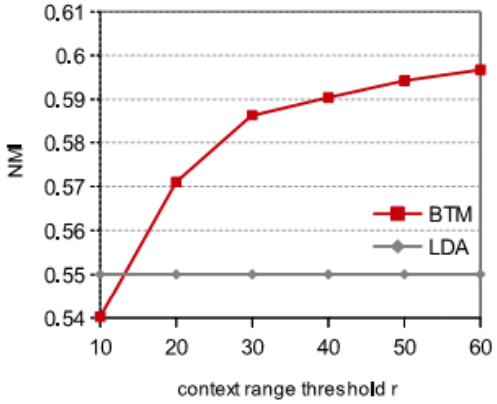


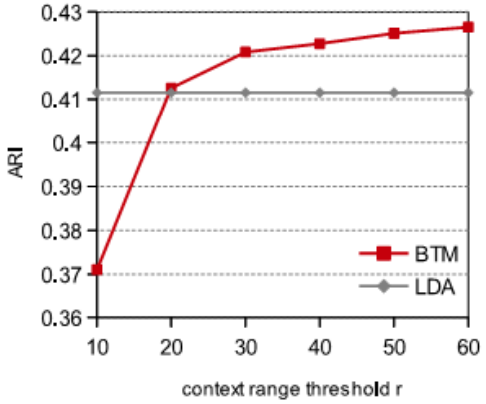


Experiments: 20Newsgroup Collection (Normal Texts)

- Biterm extraction
  - any two words co-occurring closely (with distance no larger than a threshold *r*)
- Clustering result







Summary

- A practical but not well-studied problem
  - topic modeling on short texts
  - conventional topic models suffer from the severe data sparsity when modeling the generation of short text messages
- A generative model: Biterm Topic Model
  - model word co-occurrences to uncover topics
  - fully exploit the rich global word co-occurrens
  - general and effective
- Furture works
  - better way to infer topic proportations for short text messages
  - explore BTM in real-world applications

More Information:

Http://xiaohuiyan.com

Thank You !