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

YAHOO! NEWS

WORLD NEWS »

- Syrian prime minister survives Dama bombing, six die
- Saudi-U.S. relations to withstand N oil boom
- Retailers to compensate victims of E disaster

Google AdWords

Ads related to **laptop**

Laptop
www.kelkoo.co.uk/**Laptop**
Search among thousands of deals and save mo

Donate Computers to Kids
www.maly.co.il/
100,000 Kids Need Your Support Help Us bridg

YouTube

NoSQL Database Tutorial part1 | Introduction to NoSql
上传者：Ahmad Naser
10,136 精选

O'Reilly Webcast: MongoDB Schema Design: How to Think
上传者：OreillyMedia
观看次数：18,885 次

twitter

WWW2013 @www2013ric
Science made easy: new
Newspaper editors vs th
#www2013
Expand

WWW2013 @www2013ric
The Dangers of Big Data
Expand

facebook

Marck Zuckerberg
Like This Page · Augus

Meeting with journalists from Br
:) — with Christopher Dominguc
Martinez Escalante and zaaaaaa
sco.il/9203A9

Q&A

Writing: What are some good habits to some good online sites available?
Follow • 1 Follower • Add Answer

Health and Wellness: Why is it that one still become darker despite the applicat
Follow • 1 Follower • Add Answer

Medicine and Healthcare: In what orde without oxygen?
Follow • 2 Followers • Add Answer

Booking.com

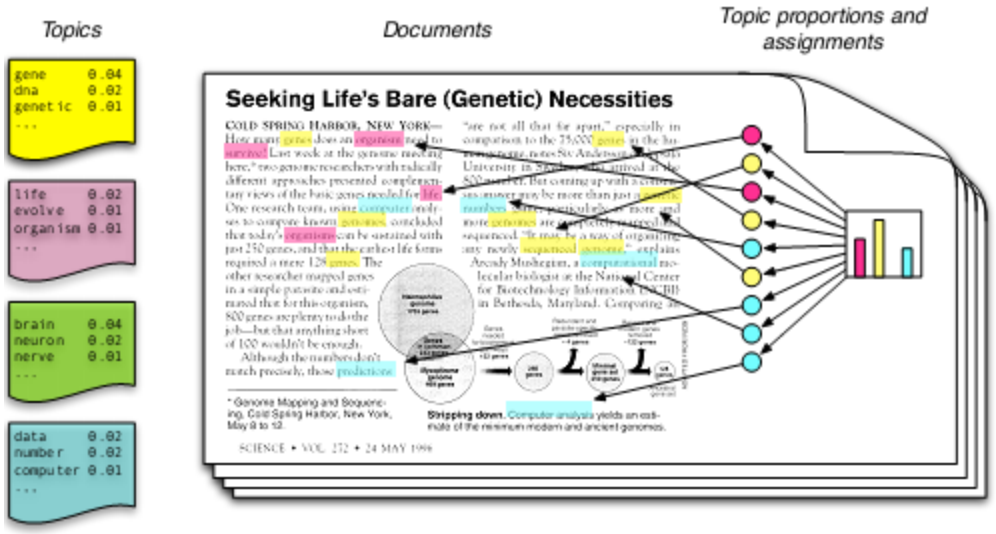
Brisa Barra Hotel ★★★★★
"The hotel was really great. We didn't want to be in ipanema or Copacaban, so we decided to go to Barra de Tijuca."
Natalia, Capital Federal

Hotel Praia Linda ★★★
"Absolutely loved it!!! Brilliant hotel! The staff are very friendly and helpful, always ready to provide the best customer service with a smile on their faces. The rooms are very clean and in good condition."
Ana, Teddington

Background

- Understanding the topics of short texts is important in many tasks
  - content characterizing
  - content recomendation
  - user interest profiling
  - emerging topic detecting
  - semantic analysis
  - ...

Topic Models



From Blei

- Model the generation of documents with latent topic structure
  - a topic ~ a probability 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 doc: most words only occur once

- Not enougth contexts to identify the senses of ambiguous words
  - normal doc: rich context, many topic-related words
  - short doc: scarce context, few topic-related words

Previous Approaches

- LDA with document aggregation
  - e.g. aggregating the tweets published by the same user
  - heuristic, not general
- Mixture of unigrams
  - each document has only one topic
  - too strict assumption, result in peaked posteriors P(z|d)
- Sparse topic models
  - each dcoument maintains a sparse distribution over topics, e.g. Focused Topic Models
  - too complex, easy to overfitting

Key Ideas

- Topics are basically groups 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?**

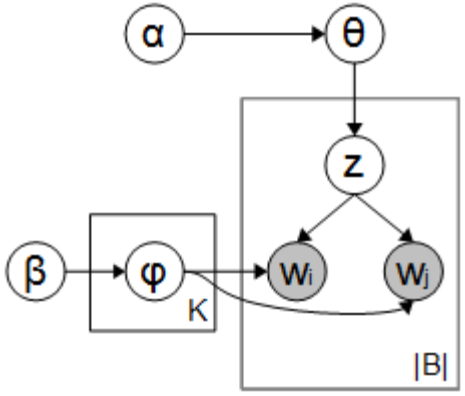
- Topic models on short texts suffer from the problem of severe sparse patterns in short documents
  - **why not use the rich global word co-occurrence patterns for better revealing topics?**

Biterm Topic Model (BTM)

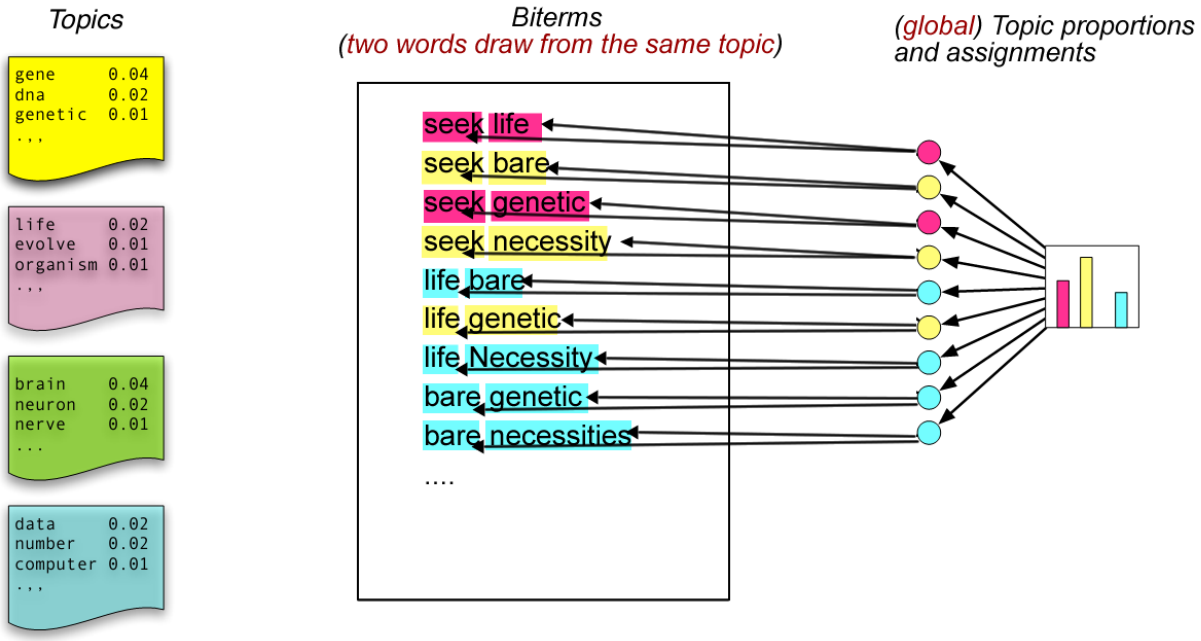
- BTM models the generation of word co-occurrences in a corpus
  - A biterm is an unordered word pair co-occurring in the same short context (document)
  - Training data includes all the biterms in the corpus

- Generative description

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



Biterm Topic Model (BTM)



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

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

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

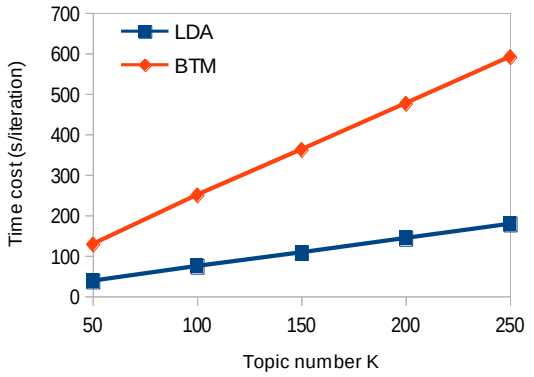
Parameters Inference

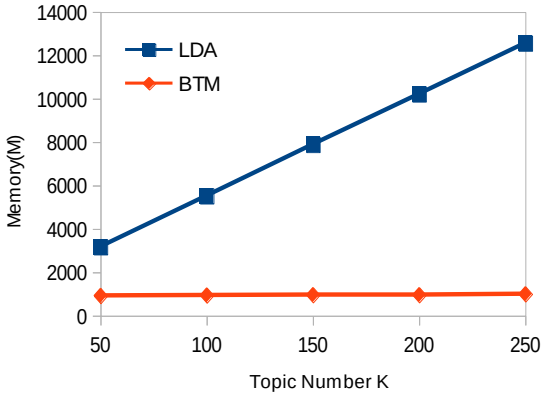
- Gibbs Sampling
  - sample topic for each biterm

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

  - parameters estimate

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





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	**5.21**	**3.94**	**97.20**

Experiments: Tweets2011 Collection

- Topic quality
  - evaluation 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

Topics selected by the word "job" on the Tweets collection. The first row lists the top 20 words, while the second row lists non-top words ranked from 1001 to 1020 based on $P(w|z)$.

LDA	LDA-U	Mixture of unigrams	BTM
job jobs business web website google design online marketing site blog project manager search www company service sales services post	job jobs design manager project web website site business service company hiring www support sales services london blog senior engineer	jobs job business marketing social media online web design website manager blog project seo internet sales tips company site hiring	jobs job manager business sales hiring service services project company senior engineer management marketing nurse office assistant center customer development
nonprofit gallery announced presence published converting select reps requirement mgr territory recruiters power involved announce poster larry dynamics feeds bristol	expertise unemployed med iii host educational fort tags apps assignments labor introduction leads github assurance avon manchester starting automotive table	understand rep industrial sustainability rankings scholarships stay single campus extra cheap 101 vp relationships beginners colorado compliance face winning mechanical	springfield mlm recruit oil req unemployment processing overview awards recruiters ict finish entrepreneur comp assist 1000 alliance locations patent auditor

The colored words are irrelevant judged by human

Experiments: Tweets2011 Collection

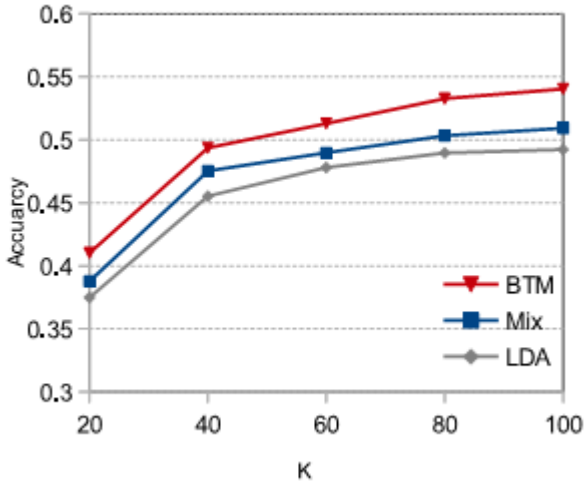
- Quality of topical representation of documents (i.e. P(z|d))
  - select 50 most frequent and meanful hashtag as class labels
  - organize documents with the same hashtag into a cluster
  - measure: H score
    - smaller value indicates better agreement with human labeled classes

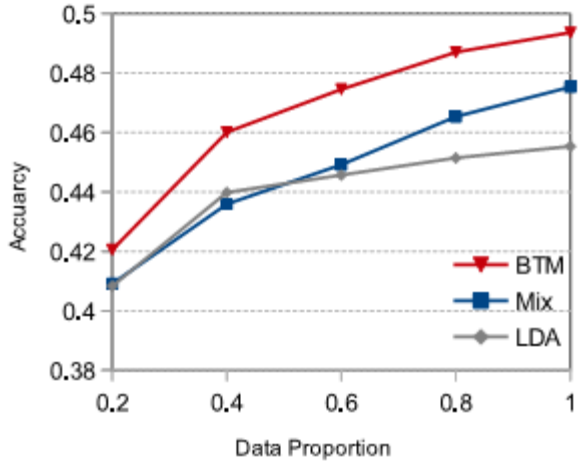
$$H = \frac{IntraDis(C)}{InterDis(C)}$$

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)

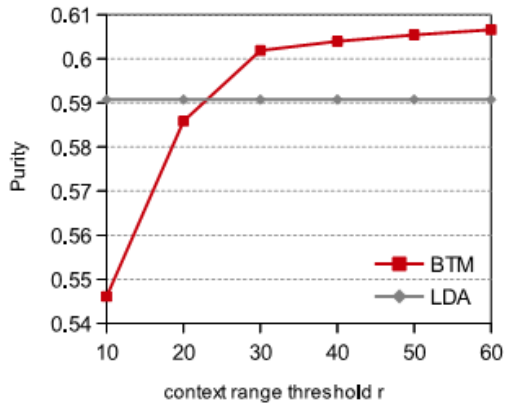


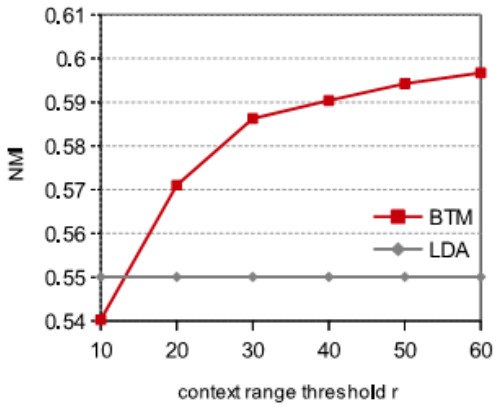


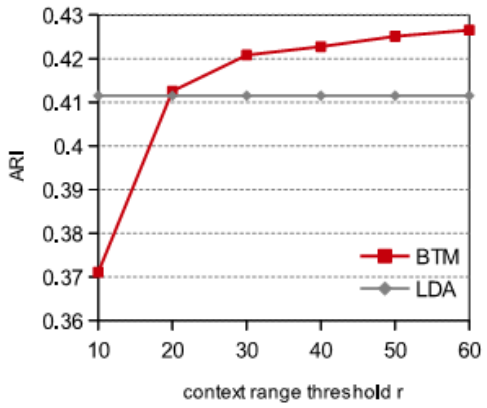
Experiments: 20Newsgroup Collection (Normal Texts)

- Biterm extraction
  - two words co-occur within a context window with range 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
- A novel way: Biterm Topic Model
  - model word co-occurrence to uncover topics
  - fully exploit the rich global word co-occurrens
  - effective on short texts (and normal texts)
- Furture works
  - better way to infer topic proportation for short text documents
  - explore BTM in real-world applications

Thank You !