

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



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

WWW2013 @www2013ric
 The Dangers of Big Data
 Expand

facebook



Marck Zuckerberg
 Like This Page · August

Meeting with journalists from Brazil :) — with Christopher Dominguez Martinez Escalante and zaaaaaa SLU.IV9ZU3AY

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 application?
 Follow · 1 Follower · Add Answer

Medicine and Healthcare: In what order 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 🇬🇧

YAHOO! NEWS

WORLD NEWS »

- Syrian prime minister survives Damascus bombing, six die
- Saudi-U.S. relations to withstand No. 1 oil boom
- Retailers to compensate victims of earthquake disaster

Google AdWords

Ads related to laptop ⓘ

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

Donate Computers to Kids
www.maly.co.il/
 100,000 Kids Need Your Support Help Us bridge the digital divide



YouTube

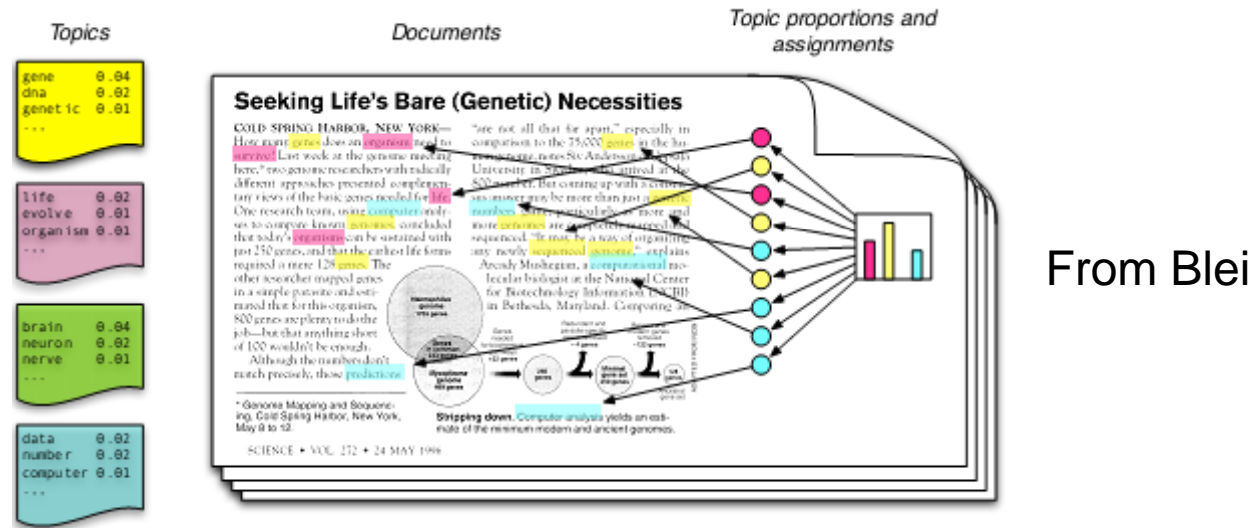
NoSQL Database Tutorial part1 | Introduction to NoSql
 上传者 : Ahmad Naser
 10,136 精选

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

Background

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

Topic Models



- 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 enough 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

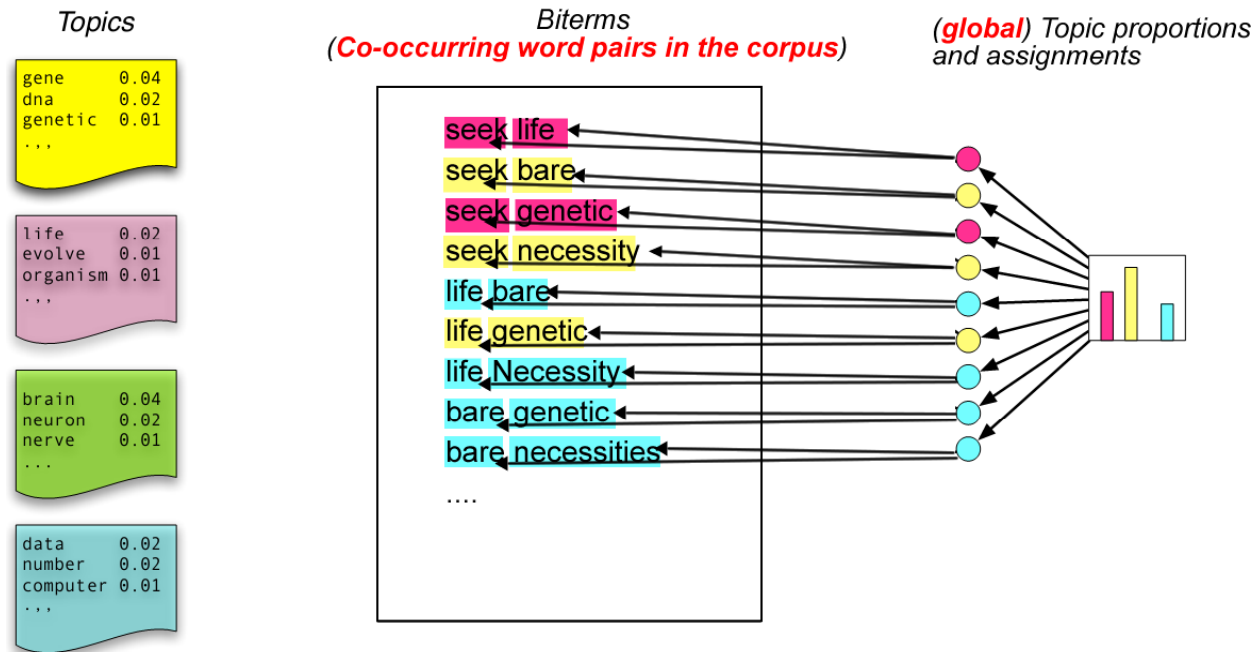
Previous Approaches on Short Texts

- 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?**

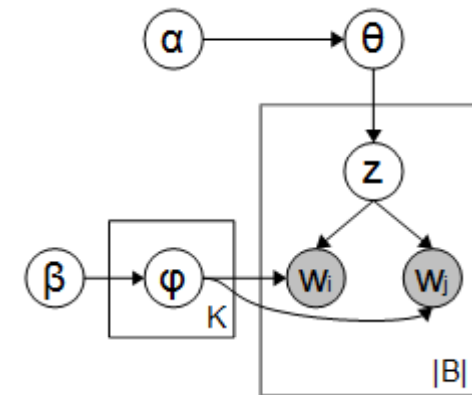
Biterm Topic Model (BTM)



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



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

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

Parameters Inference

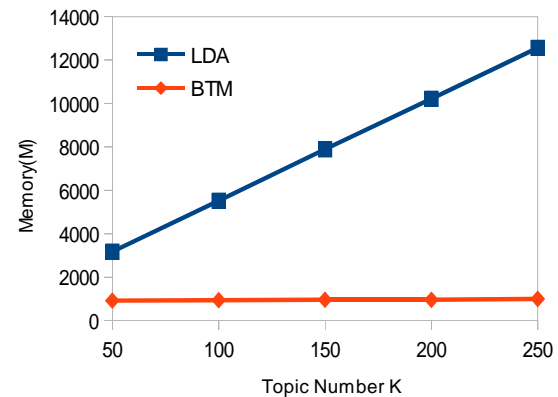
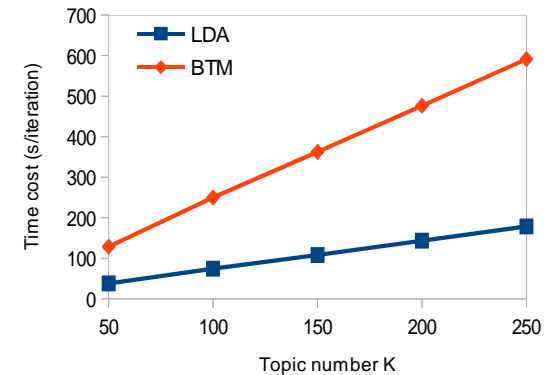
- Gibbs Sampling
 - sample topic for each biterm

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

- 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	-52.4 ± 0.1	-227.8 ± 0.3	-990.2 ± 3.8

Experiments: Tweets2011 Collection

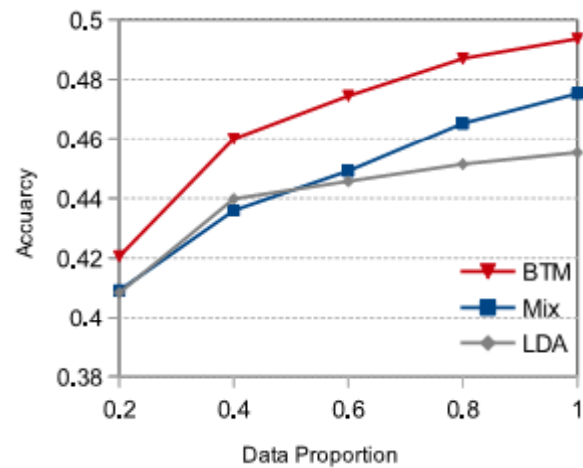
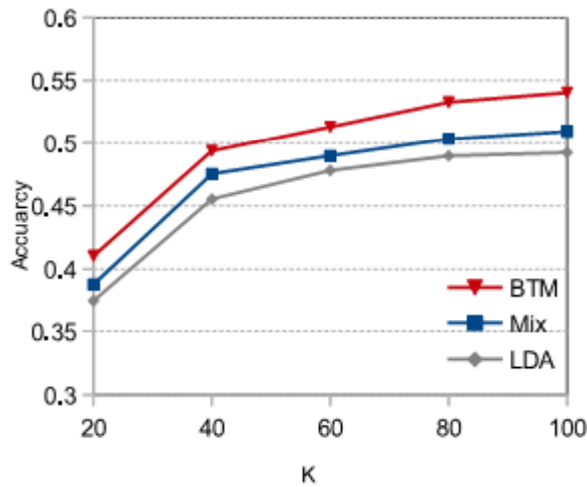
- Quality of topic proportions of documents (i.e. $P(z|d)$)
 - select 50 frequent and meaningful hashtags as class labels
 - organize documents with the same hashtag into a cluster
 - measure: H score
 - smaller value indicates better agreement with human labeled classes

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

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***

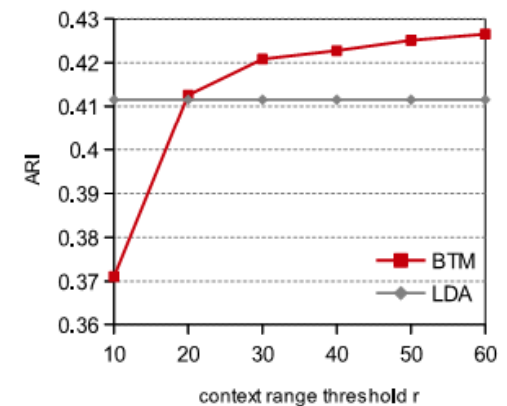
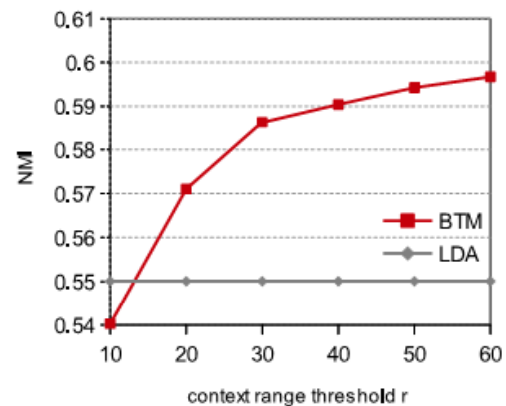
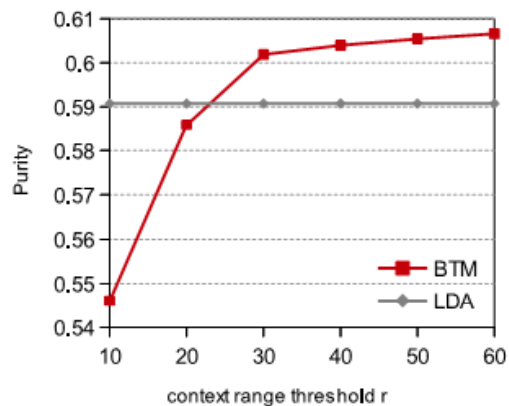
Experiments: Question Collection

- 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-occurrences
 - general and effective
- Future works
 - better way to infer topic proportions for short text messages
 - explore BTM in real-world applications

More Information:

[Http://xiaohuiyan.com](http://xiaohuiyan.com)

Thank You!

