Learning Topics in Short Texts by Non-negative Matrix Factorization on Term Correlation Matrix

Xiaohui Yan[1], Jiafeng Guo[1], Shenghua Liu[1], Xueqi Cheng[1], Yanfeng Wang[2]
[1]*Institute of Computing Technology, Chinese Academy of Sciences* , [2]*Sogou Inc.*
yanxiaohui@software.ict.ac.cn, {guojiafeng,liushenghua,cxq}@ict.ac.cn,wangyanfeng@sogou-inc.com

1. BACKGROUND & PROBLEM

1.1 Short Text

- Short texts are prevalent on the web
  - microblogs
  - SNS statuses
  - instant messages
  - webpage titles
  - advertisements
  - ...
- Distilling semantic structures in Short text is important
  - emerging topics discovery
  - efficient index and retrieval personalized
  - personalized recommendation
  - advertisement targeting
  - ...

1.2 Topic Models

- A principled way to uncover the hidden thematic structure in a large text collection
  - probabilistic models
    - PLSA, LDA, HDP, etc.
  - non-probabilistic
    - NMF (Non-negative Matrix Factorization)
- Matrix factorization view of topic models

$$\min_{U\geq 0, V\geq 0} J(U,V) = L(X,UV)$$

$X \in R^{n\times n}$ = $U \in R^{n\times k}$ × $V \in R^{k\times n}$

term-doc matrix | term-topic matrix | topic-doc matrix

1.3 Problems on Short Texts

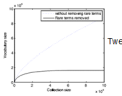
- Term-doc matrix is too sparse

Dataset	Density
Tweet	0.0034
Title	0.0038
Question	0.0012

- Making the factorization highly underdetermined
  - observed varibles much less than parameters to be estimated

1.4 Motivation

- learn topics directly from topic correlation data
  - Topics are mainly uncovered based on the correlations between terms
  - Term correlation can be estimated word co-occurrences, not necessarily dependent on document length
  - The number of distinct terms usually keeps relative small and stable



Tweet data

2. OUR APPROACH

2.1 Term Represenation

- Opt 1: document vector

$$t_i = (d_{i1}, d_{i2}, ..., d_{iN})$$

  - term weighting: tfidf
- Opt 2: co-occurred word vector

$$t_i = (w_{i1}, w_{i2}, ..., w_{iM})$$

  - term weighting: PPMI

$$w_{i,m} = \mathrm{PPMI}(t_i, t_m) = \max(\log \frac{P(t_i,t_m)}{P(t_i)P(t_m)}, 0),$$

$$P(t_i,t_m) = \frac{u(t_i,t_m)}{\sum_{j,l} u(t_j,t_l)}, P(t_i) = \frac{\sum_m u(t_i,t_m)}{\sum_{j,l} u(t_j,t_l)}$$

2.2 Term Correlation Matrix

- Term correlation is measured by cosine similiarity
  - When M<<N, the term correlation matrix computed via the co-occurred word vector representation is much denser than via the document vector represenation

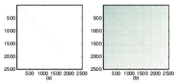


Figure 3: Visualization of Term correlation matrix on Tweets corpus computed via (a) document vector representation of terms, density=0.0415; (b) via co-occurred term vector representation of terms, density=0.8835.

2.3 Topics Learning

- After computing the term correlation matrix *S*, solving the term-topic matrix *U* by minimize

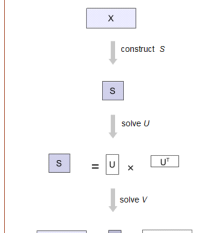
$$L(U) = \|S - UU^T\|_F^2, \quad \text{s.t. } U \geq 0$$

2.4 Topics Inference for Documents

- After *U* obtained, Solve the topic-doc matrix V by minimizing

$$L(V) = \|X - UV\|_F^2, \quad \text{s.t. } V \geq 0$$

The Overall Procedure



X: term-doc matrix S: term correlation matrix
U: term-topic matrix V: topic-doc matrix

3. EXPERIMENTS

- Datasets

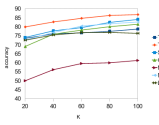
Data sets	Tweet	Title	Question
#documents	4520	2630	36219
#words	2502	1403	4956
avg words|	8.5958	5.2684	5.8092
#classes	unavailable	9	34

- Baselines
  - LDA
  - NMF
    - NMF_E: using Euclidean distance based cost function
    - NMF_I: using the generalized I-divergence
  - GNMF(graph regularized NMF)
  - SymNMF(symmetric NMF)
- TNMF: our method
  - TNMF_E: using Euclidean distance based cost function
  - TNMF_I: using the generalized I-divergence

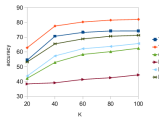
Top words in 4 topics discovered by di fferent methods on the Tweet data.

Topic	LDA	NMF_E	NMF_I	GNMF	TNMF_E	TNMF_I
Egyptian revolution	egypt food state egyptian report	egyptian egypt mubarak cairo protester	house egypt update state 100	egypt state egyptian president report	egyptian cairo protester egypt egypt	egyptian cairo protester egypt mubarak
business	service medium stand market job	market business social medium company	market red white hey die	market business social medium online	market debt credit company business	market company financial business finance
weather	hot wind change fall humidity	wind humidity temperature rain mph	snow fall wind street humidity	wind humidity rain temperature mph	humidity temperature mph hpa pressure	humidity temperature mph hpa wind
Super Bowl	super bowl green team fan	super bowl xlv sunday party	super bowl green red heart	super bowl xlv packer sunday	packer nfl bay bowl rodgers	packer nfl bay bowl rodgers

Document Classification

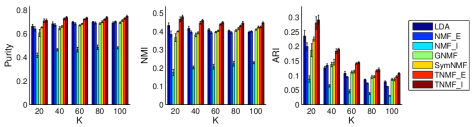


Question data

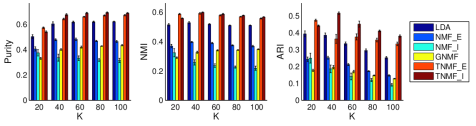


Question data

Document Clustering



Title data



Question data

What we learned: 1) Meaningful Topics can be learned from term correlations directly; 2) For short texts, learning topics from term correlation matrix is better than factorizing the term-document matrix.