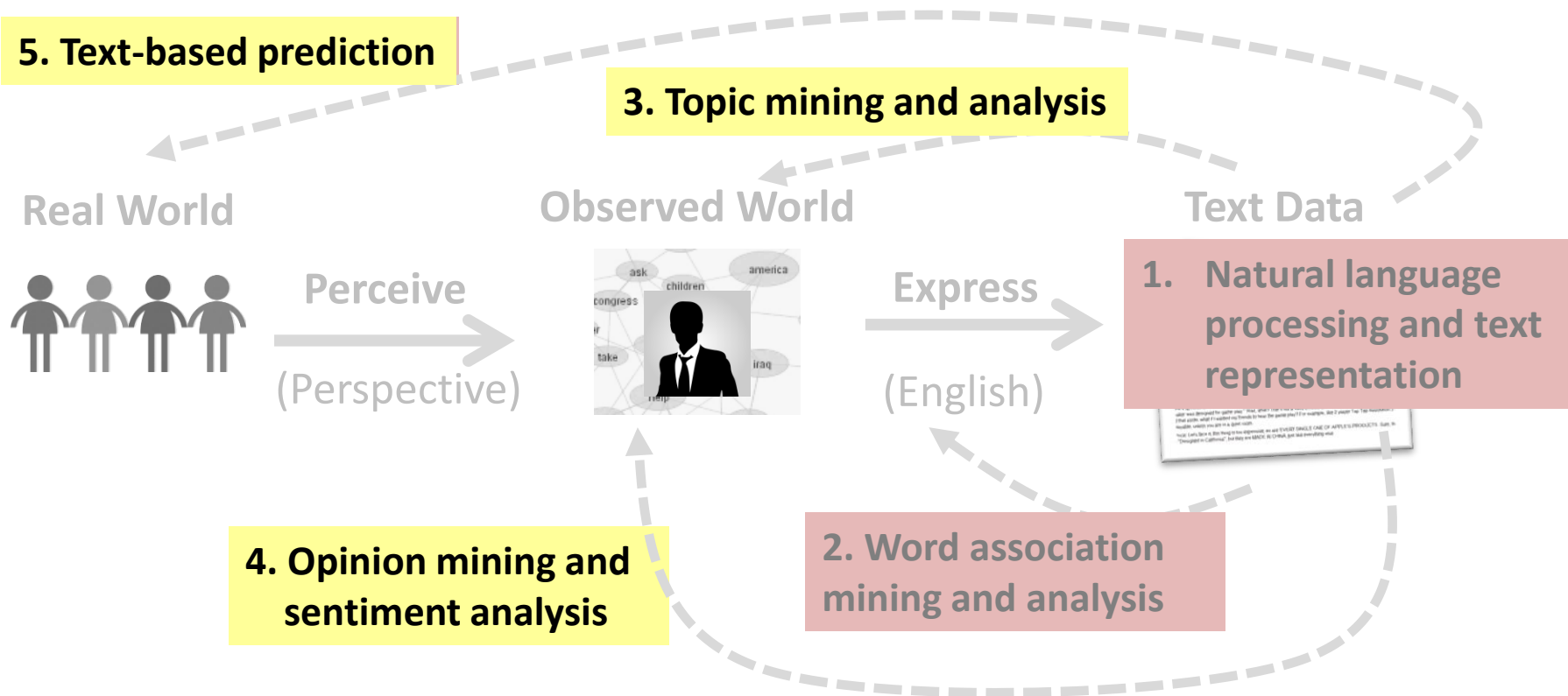




# Contextual Text Mining: Contextual Probabilistic Latent Semantic Analysis

ChengXiang “Cheng” Zhai  
Department of Computer Science  
University of Illinois at Urbana-Champaign

# Contextual Text Mining: Contextual Probabilistic Latent Semantic Analysis



# Contextual Probabilistic Latent Semantic Analysis (CPLSA)

## [Mei & Zhai 06]

- General idea:
  - Explicitly add interesting context variables into a generative model (→ enable discovery contextualized topics)
  - Context influences both coverage and content variation of topics
- As an extension of PLSA
  - Model the conditional likelihood of text given context
  - Assume context-dependent views of a topic
  - Assume context-dependent topic coverage
  - EM algorithm can still be used for parameter estimation
  - Estimated parameters naturally contain context variables, enabling contextual text mining

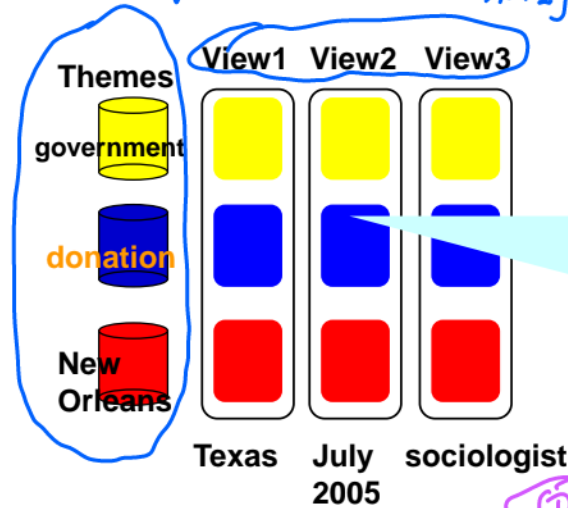
这本身就会受背景的影响

# Generation Process of CPLSA

more topics

不同的观点，并有不同的单词分布。

③ Choose a topic



government 0.3  
response 0.2..

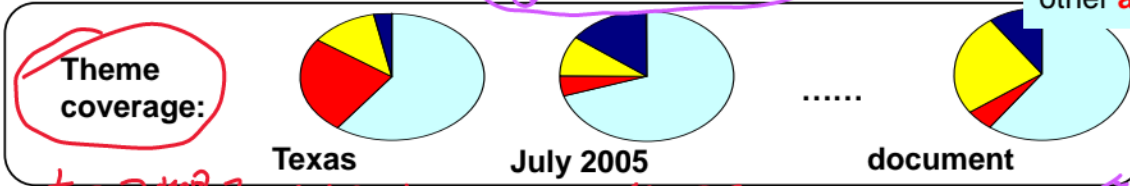
donate 0.1  
relief 0.05  
help 0.02 ..

city 0.2  
new 0.1  
orleans 0.05 ..

Draw a word f

Criticism of government response to the hurricane primarily consisted of criticism of its response to ... The total shut-in oil production from the Gulf of Mexico ... approximately 24% of the annual production and the shut-in gas production ... Over seventy countries pledged monetary donations or other assistance. ...

① Choose a view



② Choose a Coverage

在不同背景下，人们对 theme 的侧重点不同。

利用 PLSA 相同。

# Comparing News Articles [Zhai et al. 04]

## Iraq War (30 articles) vs. Afghan War (26 articles)

The common theme indicates that “United Nations” is involved in both wars

	Cluster 1	Cluster 2	Cluster 3
<b>Common Theme</b>	united 0.042 nations 0.04 ...	killed 0.035 month 0.032 deaths 0.023 ...	...
<b>Iraq Theme</b>	n 0.03 Weapons 0.024 Inspections 0.023 ...	troops 0.016 hoon 0.015 sanches 0.012 ...	...
<b>Afghan Theme</b>	Northern 0.04 alliance 0.04 kabul 0.03 taleban 0.025 aid 0.02 ...	taleban 0.026 rumsfeld 0.02 hotel 0.012 front 0.011 ...	...

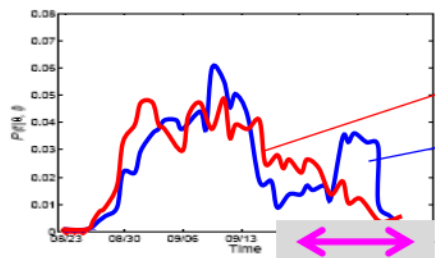
更多在  
 武器  
 层面  
 更多在  
 政治  
 层面

Collection-specific themes indicate different roles of “United Nations” in the two wars

# Theme Life Cycles in Blog Articles About “Hurricane Katrina” [Mei et al. 06]

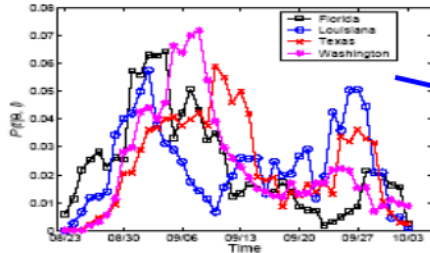
时间作为  
增量

两个话题  
的时间  
趋势。



(a) Theme life cycles in Texas (Hurricane Katrina)

**Hurricane Rita**



(b) Theme “New Orleans” over states (Hurricane Katrina)

*Oil Price*

*New Orleans*

*price 0.0772*  
*oil 0.0643*  
*gas 0.0454*  
*increase 0.0210*  
*product 0.0203*  
*fuel 0.0188*  
*company 0.0182*  
...

*city 0.0634*  
*orleans 0.0541*  
*new 0.0342*  
*louisiana 0.0235*  
*flood 0.0227*  
*evacuate 0.0211*  
*storm 0.0177*  
...

# Spatial Distribution of the Topic “Government Response” in Blog Articles About Hurricane Katrina

位置作为背景 [Mei et al. 06]



(a) Week 1: 08/23-08/29



(b) Week Two: 08/30-09/05



(c) Week Three: 09/06-09/12

Theme 1
Government Response
bush 0.0716374
president 0.0610942
federal 0.0514114
govern 0.0476977
lema 0.0474692
administrate 0.0233903
response 0.0208351
brown 0.0199573
blame 0.0170933
governor 0.0142153



(d) Week Four: 09/13-09/19



(e) Week Five: 09/20-09/26

# Event Impact Analysis: IR Research [Mei & Zhai 06]

Topic: retrieval models

<i>term</i>	0.1599
<i>relevance</i>	0.0752
<i>weight</i>	0.0660
<i>feedback</i>	0.0372
<i>independence</i>	0.0311
<i>model</i>	0.0310
<i>frequent</i>	0.0233
<i>probabilistic</i>	0.0188
<i>document</i>	0.0173
...	

<i>vector</i>	0.0514
<i>concept</i>	0.0298
<i>extend</i>	0.0297
<i>model</i>	0.0291
<i>space</i>	0.0236
<i>boolean</i>	0.0151
<i>function</i>	0.0123
<i>feedback</i>	0.0077
...	

<i>xml</i>	0.0678
<i>email</i>	0.0197
<i>model</i>	0.0191
<i>collect</i>	0.0187
<i>judgment</i>	0.0102
<i>rank</i>	0.0097
<i>subtopic</i>	0.0079
...	

SIGIR papers

A seminal paper [Croft & Ponte 98]

1992

Star

用重大事件  
作为锚点

<i>probabilist</i>	0.0778
<i>model</i>	0.0432
<i>logic</i>	0.0404
<i>ir</i>	0.0338
<i>boolean</i>	0.0281
<i>algebra</i>	0.0200
<i>estimate</i>	0.0119
<i>weight</i>	0.0111
...	

1998

<i>model</i>	0.1687
<i>language</i>	0.0753
<i>estimate</i>	0.0520
<i>parameter</i>	0.0281
<i>distribution</i>	0.0268
<i>probable</i>	0.0205
<i>smooth</i>	0.0198
<i>markov</i>	0.0137
<i>likelihood</i>	0.0059
...	

year

时间  
作为  
锚点



# Suggested Reading

- **[Zhai et al. 04]** ChengXiang Zhai, Atulya Velivelli, and Bei Yu. 2004. A cross-collection mixture model for comparative text mining. In *Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining* (KDD 2004). ACM, New York, NY, USA, 743-748. DOI=10.1145/1014052.1014150
- **[Mei & Zhai 06]** Qiaozhu Mei and ChengXiang Zhai. 2006. A mixture model for contextual text mining. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining* (KDD 2006). ACM, New York, NY, USA, 649-655. DOI=10.1145/1150402.1150482
- **[Mei et al. 06]** Qiaozhu Mei, Chao Liu, Hang Su, and ChengXiang Zhai. 2006. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of the 15th international conference on World Wide Web* (WWW 2006). ACM, New York, NY, USA, 533-542. DOI=10.1145/1135777.1135857