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

5. Text-based prediction

3. Topic mining and analysis

Real World

Observed World

Text Data

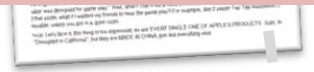
Perceive

(Perspective)

Express

(English)

1. **Natural language processing and text representation**



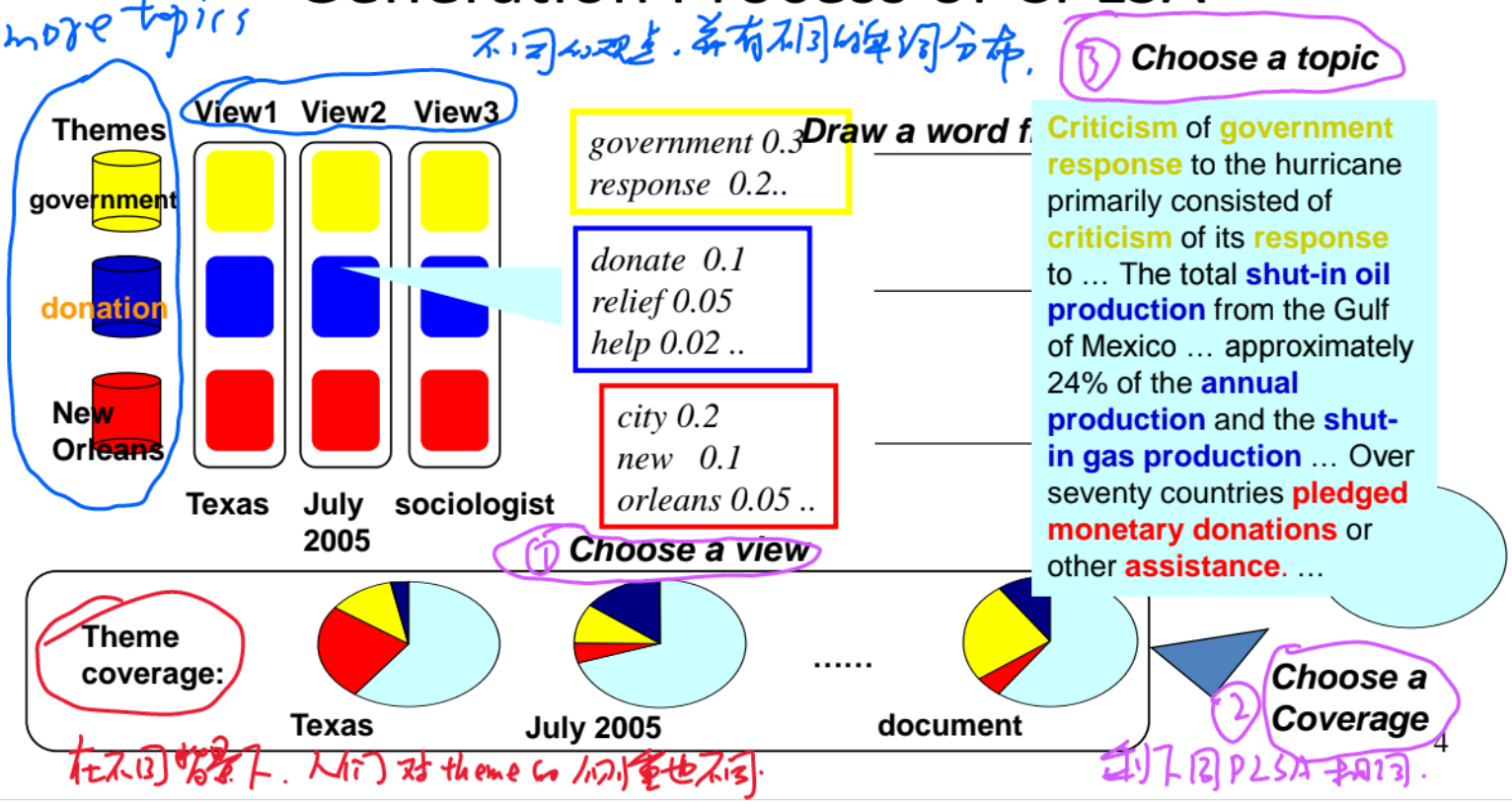
2. Word association mining and analysis

4. Opinion mining and sentiment 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

Generation Process of CPLSA



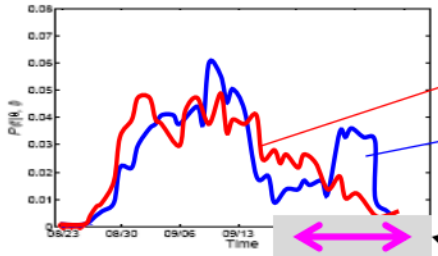
Comparing News Articles [Zhai et al. 04]
Iraq War (30 articles) vs. Afghan War (26 articles)

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

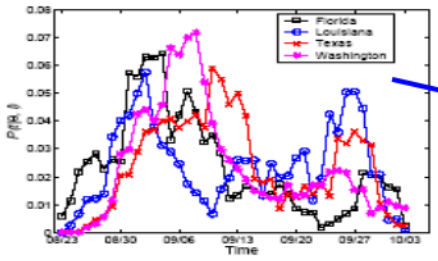
	Cluster 1	Cluster 2	Cluster 3
Common Theme	united 0.042 nations 0.04 …	killed 0.035 month 0.032 deaths 0.023 …	…
Iraq Theme	n 0.03 Weapons 0.024 Inspections 0.023 …	troops 0.016 hoon 0.015 sanches 0.012 …	…
Afghan Theme	Northern 0.04 alliance 0.04 kabul 0.03 taleban 0.025 aid 0.02 …	taleban 0.026 rumsfeld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

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]



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



(b) Theme "New Orleans" over states (Hurricane Katrina)

Hurricane Rita

Oil Price

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

New Orleans

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 [Mei et al. 06]

(a) Week1: 08/23-08/29

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

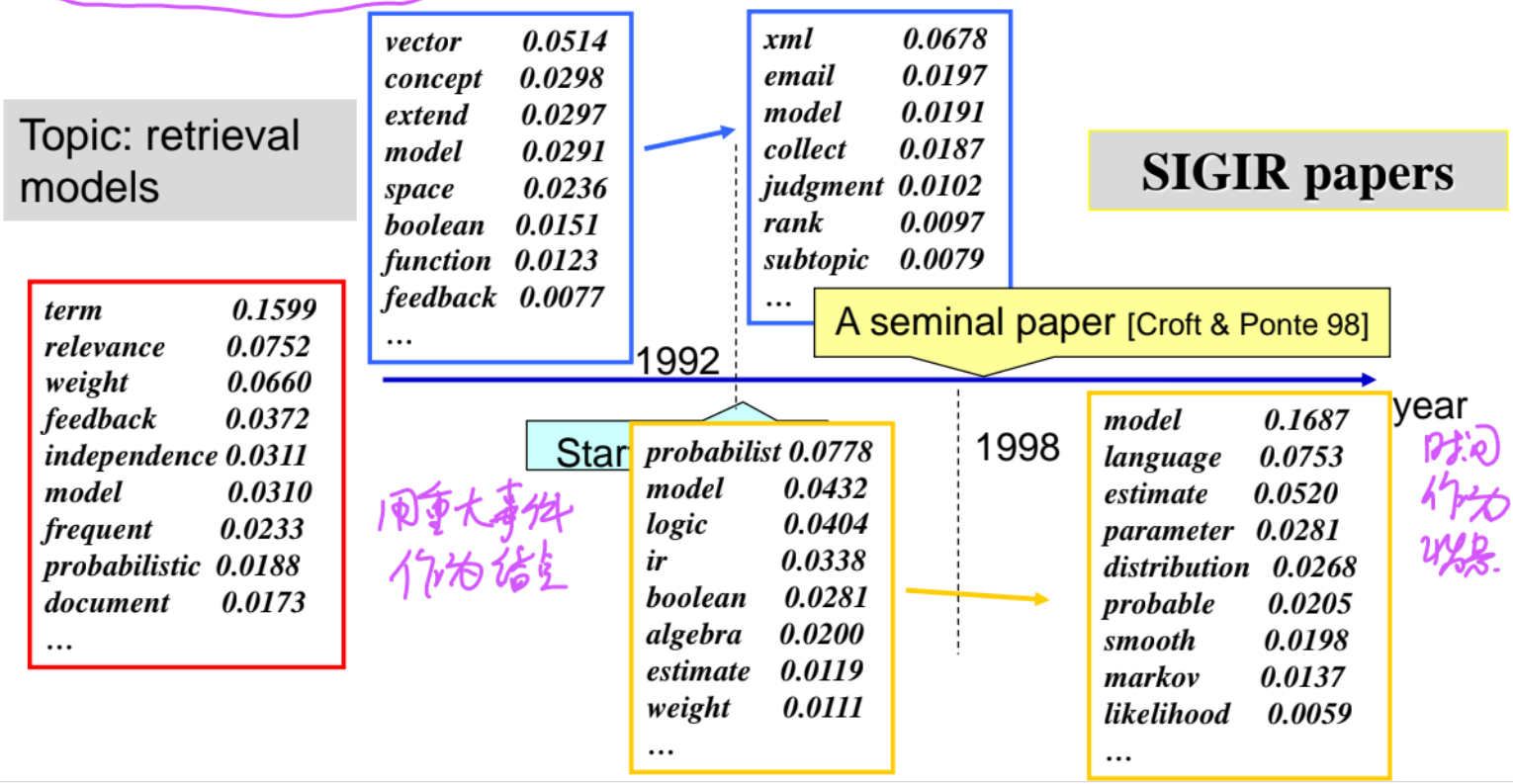
(c) Week Three:09/06-09/12

Theme 1
Government Response
bush 0.0716374
president 0.0610942
federal 0.0514114
govern 0.0476977
fema 0.0474692
administrate 0.0233903
response 0.0208351
brown 0.0199573
blame 0.0170033
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]



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