Contextual Text Mining with Probabilistic Topic Models

Qiaozhu Mei, ChengXiang Zhai

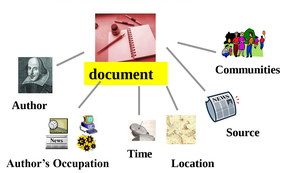
Multimodal Information Access and Synthesis Center, Department of Computer Science,
University of Illinois at Urbana-Champaign, Urbana, Illinois 61801

Introduction

The explosive growth of information demands powerful text mining tools to help us digest information and discover hidden knowledge in text. Text analysis is often associated with various kinds of context, such as time, location, sources. Given any text data with context information, we often would like to extract the subtopics or themes from text and analyze their variations over context, e.g., to reveal spatiotemporal variations of a subtopic like "government response" in blog articles about hurricane Katrina. In this project, we are developing general probabilistic models and new algorithms for discovering and analyzing various contextual patterns from text, which we refer to as *contextual text mining*. The proposed models have broad applications in multiple domains to help understand topic evolutions, spatiotemporal impact of events, public opinions, and detect topic related social communities in arbitrary text collections. The extracted topics patterns can reveal hidden associations and latent knowledge in text and provide evidence for decision-makers to use in making policy decisions.

Materials and methods

Contextualized Text



- **Many genres of massive text data**
  - news, blogs, literature, emails,…
- **Texts are associated with context**
  - time, geographic location, author, source, sentiments, social network, …
- **Many patterns rely on the analysis of context**
  - compare the variations of topics in text over different contexts
- **Context modeling is powerful**
  - discover many different patterns

Contextual Text Mining

- **Discover hidden topics/themes from text**
  - What did people say about hurricane Katrina in blogs, in news, …?
- **Reveal topic variations over contexts**
  - How do opinions about government response in hurricane Katrina vary over different states?
- **Reveal correlations of topical patterns and context variables**
  - Have positive opinions about Iraq war in blog articles been affected by special V programs Covering the war?

Key Technologies

- **General contextual text mining models**
  - Contextual probabilistic latent semantics analysis model
- **General probabilistic topic labeling**
  - Maximizing label-topic mutual information
- **Information retrieval with language models**
  - Probabilistic text representation and matching
- **Context specific techniques**
  - E.g., social network analysis

Utilities

- **Topic (theme) extraction**
- **Opinion summarization**
- **Topic trend analysis**
- **Spatiotemporal patterns**
- **Event impact analysis**
- **Communities finding**
- **Sentiment analysis**
- **…**

Results

Comparing news articles (Zhai et al. 04)

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

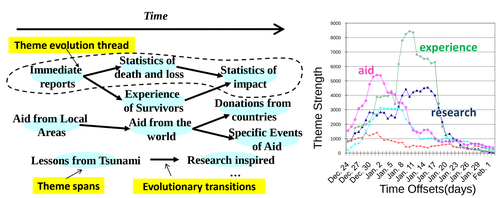
The common theme indicates that "United Nations" is involved in both wars

	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 evolution and theme life cycle in news articles (Mei et al. 05)

- News articles about the Asian Tsunami, 2005



Theme evolutions about Tsunami **Themes life cycles of CNN news**

Spatiotemporal patterns in blog articles (Mei et al. 06)

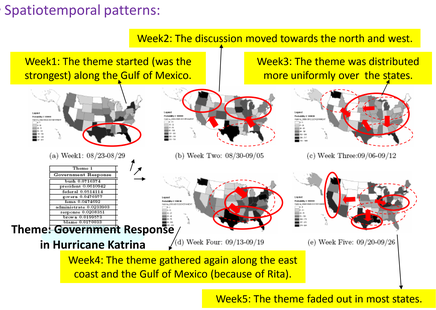
- Query= "Hurricane Katrina"

- Blogs about Hurricane Katrina, 2006

- Topics in the results:

Government Response	New Orleans	Oil Price	Praying and Blessing	Aid and Donation
bush 0.071	city 0.063	price 0.077	god 0.141	donate 0.120
president 0.061	orleans 0.054	oil 0.064	pray 0.047	relief 0.076
federal 0.051	new 0.034	gas 0.045	prayer 0.041	red 0.070
government 0.047	louisiana 0.023	increase 0.020	love 0.030	cross 0.065
fema 0.047	flood 0.022	product 0.020	life 0.025	help 0.050
administrate 0.023	evacuate 0.021	fuel 0.018	bless 0.025	victim 0.036
response 0.020	storm 0.017	company 0.018	lord 0.017	organize 0.022
brown 0.019	resident 0.016	energy 0.017	jesus 0.016	effort 0.020
blame 0.017	center 0.016	market 0.016	will 0.013	fund 0.019
governor 0.014	rescue 0.012	gasoline 0.012	faith 0.012	volunteer 0.019

- Spatiotemporal patterns:



Results (Cont.)

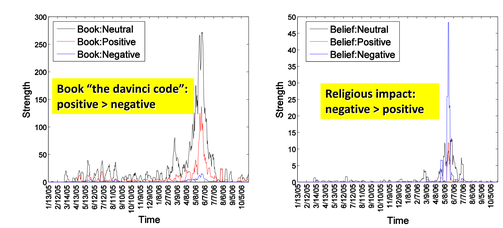
Structured summary of topics and sentiments (Mei et al. 07)

- Blogs articles about "the Da Vinci Code", 2006

	Neutral	Positive	Negative
Topic 1: Movie	… Ron Howards selection of Tom Hanks to play Robert Langdon.	Tom Hanks stars in the movie,who can be mad at that?	But the movie might get delayed, and even killed off if he loses.
	Directed by: Ron Howard Writing credits: Akiva Goldsman …	Tom Hanks, who is my favorite movie star act the leading role.	protesting … will lose your faith by … watching the movie.
	After watching the movie I went online and some research on …	Anybody is interested in it?	… so sick of people making such a big deal about a FICTION **book** and **movie**.
Topic 2: Book	I remembered when i first read the book, I finished the book in two days.	Awesome book.	… so sick of people making such a big deal about a FICTION **book** and **movie**.
	I'm reading "Da Vinci Code" now. …	So still a good book to past time.	This controversy book cause lots conflict in west society.

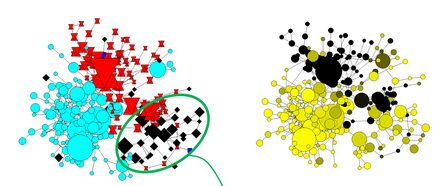
Sentiment dynamics and bursts (Mei et al. 07)

- Blogs articles about "the Da Vinci Code", 2006



Social networks and topics (Mei et al. 08)

- Bibliography and coauthor network, 2007

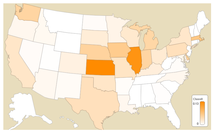


Who's in the same community, and why? (What's the semantics of a community ?) **How is a topic distributed in a network? Who belongs to multiple communities?**

Semantics of this community: "Data Mining"

Mining	0.11
Data	0.06
Discovery	0.03
Databases	0.02
Rules	0.02
Association	0.02
Patterns	0.02
Frequent	0.01
Streams	0.01

-Blog articles about the weather, 2007
- Geographic network: the Windy states



Potential applications

We plan to further explore more robust and effective contextual text mining techniques, and apply them to different real world text information management tasks. Text mining is an important topic in the "*Advanced Data Analysis and Visualization*" area. Methods developed in this project have potential applications in other areas such as:

Risk and Decision Sciences
- monitoring impact of events and policies;
- monitoring topic trends and predicting bursting topics and patterns;

Social, Behavioral and Economic Sciences
- summarizing public opinions;
- identifying social communities and their focused topics;
- monitoring the diffusion of opinions, rumors, on social network

Natural Disasters and Related Geophysical Studies
- summarizing impact of events;
- analyzing distribution and change of impact and opinion geographically.

Literature cited

- Qiaozhu Mei, Deng Cai, Duo Zhang, ChengXiang Zhai. **Topic Modeling with Network Regularization**, Proceedings of the 17th International World Wide Web Conference (**WWW' 08**). To appear.
- Qiaozhu Mei, Xuehua Shen, ChengXiang Zhai. **Automatic Labeling of Multinomial Topic Models**, Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (**KDD' 07**), pages 490-499, 2007 **Runner-up Best Student Paper Award**
- Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, ChengXiang Zhai. **Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs**, Proceedings of the 16th International World Wide Web Conference (**WWW' 07**), pages 171-180, 2007.
- Qiaozhu Mei, Dong Xin, Hong Cheng, Jiawei Han, ChengXiang Zhai. **Generating Semantic Annotations for Frequent Patterns with Context Analysis**, Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (**KDD' 06**), pages 337-346, 2006. **Runner-up Best Student Paper Award**
- Qiaozhu Mei, Chao Liu, Hang Su, and ChengXiang Zhai, **A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs**, Proceedings of the 15th International World Wide Web Conference (**WWW'06**), pages 533-542, 2006.
- Qiaozhu Mei, ChengXiang Zhai, **Discovering Evolutionary Theme Patterns from Text - An Exploration of Temporal Text Mining**, Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (**KDD' 05**), pages 198-207, 2005
- ChengXiang Zhai, Atulya Velivelli, Bei Yu, **A cross-collection mixture model for comparative text mining**, *Proceedings of ACM KDD 2004* (**KDD'04**), pages 743-748, 2004.

Acknowledgements

This project was in part funded through MIAS, the DHS University Center of Excellence by a grant from the Department of Homeland Security, Science and Technology Directorate, Office of University Programs.

For further information

For more information, please contact Qiaozhu Mei at qmei2@uiuc.edu, or visit our website at http://sifaka.cs.uiuc.edu/~qmei2/ctm.html.