

Contextual Text Mining with Probabilistic Topic Models

Qiaozhu Mei and ChengXiang Zhai
University of Illinois at Urbana-Champaign, COE
Multimodal Information Access and Synthesis Center
(Dan Roth, Principle Investigator)

Project Scope: 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, and sources. Given any text data with context information, we often would like to extract selected subtopics or themes and analyze their variations over context such as, for example, revealing 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 topic patterns can reveal hidden associations and latent knowledge in text, and provide evidence for policy-makers to use in making decisions.

Recent Progress: Probabilistic topic models have been developed to model and analyze temporal variations of topics, spatiotemporal patterns of topics, mixture of topics and sentiments, and correlations of topics and social networks.

Future Plans: We plan to further explore more robust and effective contextual text-mining techniques, and apply them to different real-world text-information management tasks. We plan to leverage the generality and effectiveness of our methods to develop novel application systems to help people access, mine, and summarize information in text data.

Relevance to listed research areas: Text mining is an important topic in the “*Advanced Data Analysis and Visualization*” area; methods developed in this project have applications in other areas such as “*Risk and Decision Sciences*,” “*Social, Behavioral and Economic Sciences*” and “*Natural Disasters and Related Geophysical Studies*.”

Publications:

1. Qiaozhu Mei, Deng Cai, Duo Zhang, ChengXiang Zhai. Topic Modeling with Network Regularization, *Proceedings of WWW 2008*. **In press**.
2. Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai, Automatic Labeling of Multinomial Topic Models, *Proceedings of ACM KDD 2007*, pages 490-499. **(Runner-up for best student paper)**
3. Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, ChengXiang Zhai, Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs, *Proceedings of WWW 2007*, pages 171-180.