**User Study for Interactive Topic Modeling**

Yuening Hu, Alison Smith and Jordan Boyd-Graber

ynhu@cs.umd.edu, amsmit@cs.umd.edu, jbg@umiacs.umd.edu

University of Maryland, College Park

---

**Introduction**

- Topic models are widely applied to discover thematic topics
- The discovered topics do not always make sense to end users
- Confuse two or more themes into one topic
- Two different topics can be (near) duplicates
- Some topics make no sense at all

We proposed Interactive Topic Modeling (ITM, Hu et al. 2013)

- Uses tree-based topic models
- Allows users to encode their feedback iteratively
- Question: Does ITM help users? and how?
- Solution: Perform user study for ITM

---

**Tree-based Topic Models**

- Model the correlations of words
- Positive correlations: encourage words to appear in the same topic
- Negative correlations: push words away from the same topic
- Use tree prior instead of a symmetric prior to model the correlations

---

**Generative process**

- Draw multinormal distributions

---

**Overview**

- Interface for ITM user study
  - Web-based application with a HTML and jQuery front end, connected via Ajax and JSON
  - Includes two parts: refine topics and a test
  - Multiple ways to explore the corpus to answer the questions

- Start page and topic display

---

**Query and answer questions**

**Interactive Topic Modeling**

- Encode users' feedback as correlations
- Give non machine learning experts a voice to update topics
- Update the topics interactively and iteratively

---

**User Study**

- Evaluate whether and how ITM helps users to understand data
- User study set up
  - Refining topics (15 mins) and a test (30 mins) (Wacholder and Liu, 2008)
  - Experimental group (ITM) and control group (LDA)
  - 10 questions including legislation about taxes, US-Mexico border, etc
  - 10 topics, 20 users (randomly assigned to a group)

- Legislative corpus
  - Transcripts of political debates for the 109th congress (2005 to 2006)
  - Available online from GovTrack
  - A subset of legislative bills on immigration, estate tax, stem cell research, etc
  - Use each turn as a document for topic modeling, so 2250 documents

- User population
  - 20 participants (10 for each group), all fluent in English
  - Participants are either students pursuing a degree in computer science, information science, linguistics, or working in a related field
  - Most users have little or no knowledge about congressional debates (a post-test user survey)
  - Users have varied experience with topic models

- Evaluating the refined topics
  - Each turn (document) is associated with a single bill, the true cluster label
  - Assign each doc to a cluster of its highest-probability topic
  - Variation of information: information distance between two partitions
  - "x2" found: homeland security, immigration, abortion, energy, flag burning, etc
  - "x2" successfully reduced variation of information (so did the other users)

- More analysis
  - Some users in control group performed well and used the initial topics a lot
  - Some users in experimental group complained they didn't have enough time
  - Most users like the interface
  - Users from both groups said that topics helped them to answer questions
  - Some users commented some questions were too detailed
  - No significant difference between the scores of two groups

- Graded by two graders and compared the scores of users in two groups
  - "x2" successfully reduced variation of information (so did the other users)

---

**Conclusion**

- ITM provides a tool for non-machine-learning experts to update topics
- ITM assists users in exploring a large corpus
- Topic modeling is helpful for users to understand documents
- Need more users: our population was too diverse and too small

---

Email: ynhu@cs.umd.edu, jbg@umiacs.umd.edu

Webpage: http://www.cs.umd.edu/~ynhu/