VISUALIZING SEMANTIC SPACE OF ONLINE DISCOURSE: THE KNOWLEDGE FORUM CASE

UNIVERSITY OF TORONTO
OISE | ONTARIO INSTITUTE FOR STUDIES IN EDUCATION

BODONG CHEN, UNIVERSITY OF TORONTO [bodong.chen@gmail.com]

BACKGROUND

Online discussion has been widely applied to support collaborative learning. As its usage has grown, so has the need for effective tools to interpret and facilitate online discussion. Analysis of textual information has maintained to be a focus in this research area. In particular, a rich body of literature has studied ways to combine text mining with visualization techniques to interpret and represent online discussion to promote reflection and "metadiscourse."

Research of **knowledge building** (KB) has maintained a long-standing interest in analyzing online discussion given its emphasis on idea improvement through communal **discourse**. Analytics of KB discourse has been mainly concerned with the evolution of community knowledge. An interesting challenge in KB research is to analytically tackle the following questions: "What is the state of community knowledge?", "Is the discourse effective in advancing knowledge?", and "Where is the community discourse headed?"

GOALS

Building on prior work, the present study aims to apply text mining and visualization techniques to achieve the following goals:

1. To model semantic space of KB discourse focusing on its emergent topics
2. To visualize interpreted semantic space and explore roles of different variables in KB discourse
3. To track student participation in different discussion topics in order to assess development of knowledge

METHOD

The following techniques are proposed to achieve these goals:

1. **Latent Dirichlet Allocation (LDA)**—to model topics in KB discourse and subsequently convert it into a high-dimensional semantic space
2. **Locally Linear Embedding (LLE)**—to reduce dimensions of the modeled semantic space for 2D visualization
3. **Time Series Analysis**—to track development of individual topics and to assess their development over time

A well-documented Knowledge Forum dataset from a Grade 4 science unit, "Light", was used for a case study. It contains 308 *notes* from six *views*, including: "How Light Travels", "Light and Materials", "Natural and Artificial Light", "Colours of Light", "Shadows", and "All We See Is Light?" Through content analysis, previous research has identified 15 major discussion topics from this dataset (Zhang et al., 2007).

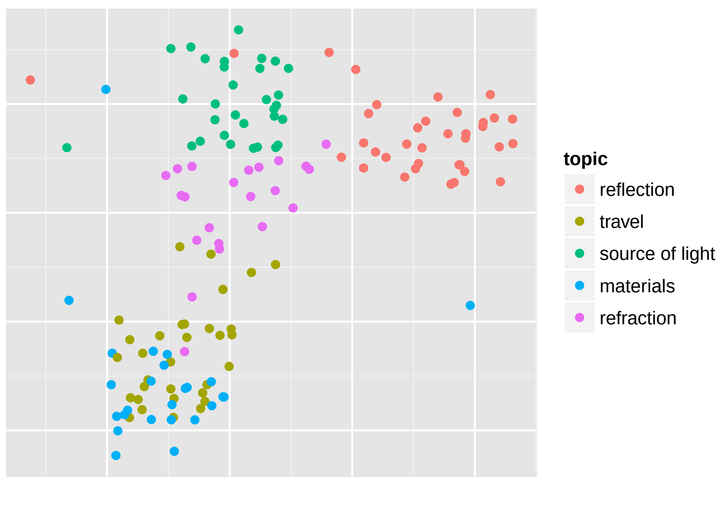
STEP 1: MODEL SEMANTIC SPACE WITH LDA

Based on analysis of the dataset in previous research, I first trained an LDA model with 15 *topics*. For each *note*, its probability of belonging to each topic were computed. Each topic is associated with a list of *terms*. By analyzing terms, I assigned a name to each topic and found them consistent with previous results from content analysis. The top five topics, which covered 150 of all 308 notes, are presented below.

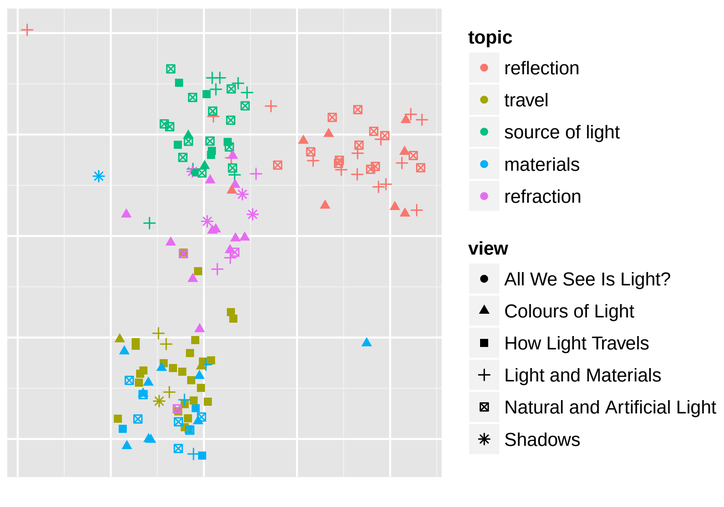
Terms	Number of notes	Topic name
mirror hot look angl bounc	36	Reflection
rainbow make prism thing white	33	Refraction
glow worm chemic look thing	29	Sources of light
travel wave line straight becaus	27	How light travels
water paper made part turn	25	Light and materials

STEP 2: VISUALIZE DIMENSION-REDUCED SEMANTIC SPACE

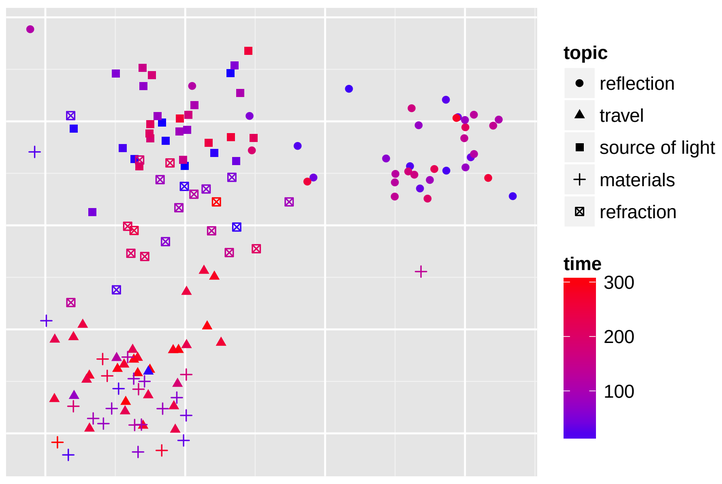
Using LLE, the 15-dimensional space was reduced to 2D. Notes were then visualized by different variables—topics, views, time, and students—to explore usefulness of the analysis.



(1) By topics (top 5 topics)



(2) By topics and views

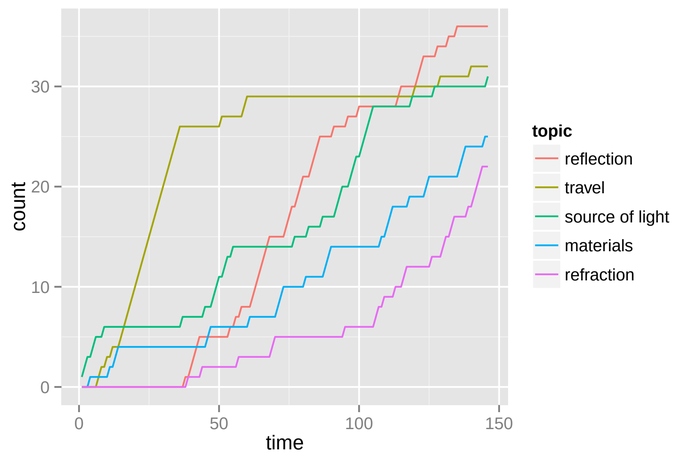


(3) By time

(4) By students (top 3)

STEP 3: TRACE TOPICS

Applying time series analysis, participation in the top five topics is visualized below.



According to this figure, different topics had their "rises and falls." For example, the beginning of the discourse focused on "how light travels," while "refraction" and "reflection" dominated the end. "Sources of light" and "light and materials" were discussed throughout the unit, while the other three topics had their own "golden times."

FUTURE DIRECTIONS

Knowledge Forum is currently under re-development and learning analytics is treated as an integral part of its functionalities. Powerful techniques are needed to turn discourse into insights to be fed back, or "forward," to boost discourse. The presented study represents an early experimentation on this front. For next steps, I will test this approach with richer KF data and fine tune it for more meaningful and usable results. I am also planning to expand this work to other learning contexts such as Massive Open Online Courses (MOOCs).

SOURCE CODE

The source code is available at https://github.com/dirkchen/topicmodels