Analyzing Online Knowledge Building Discourse Using Probabilistic Topic Models

Weiyi Sun, Department of Informatics, University at Albany
Jianwei Zhang, Hui Jin, School of Education, University at Albany
Siwei Lyu, Department of Computer Science, University at Albany
1400 Washington Ave, Albany NY 12222
Email: wsun2@albany.edu, jzhang1@albany.edu, nmjinhui@163.com, lsw@cs.albany.edu

Abstract: This exploratory study tested the use of machine learning techniques, in particular, probabilistic topic models, to conduct automated analysis of the online discourse a Grade 4 knowledge-building community that investigated optics over three months. Using the Latent Dirchilet Allocation (LDA) model, we extracted ten distinct and semantically meaningful clusters (i.e., topics) from the online discourse, which overlapped substantially with — although did not directly map onto—the inquiry themes identified by students and inquiry thread topics identified by researchers. The LDA analysis further identified discourse entries relevant to each of the topics, with a high-level agreement achieved between the automated analysis results and the manual coding of two researchers.

Introduction

A focal challenge in the learning sciences community is to create collaborative knowledge building environments in line with how real-world knowledge communities work. Supported by collaborative online environments, such as Knowledge Forum (Scardamalia & Bereiter, 2006), students engage in sustained and interactive discourse to advance their collective understanding. Working as a knowledge building community requires students to take on collective responsibility for advancing their collective knowledge beyond individual learning (Scardamalia & Bereiter, 2006; Stahl, 2006). They need to develop a reflective awareness of the major themes of inquiry emerging from the diverse input of all community members, build on important ideas of peers, identify progress and gaps, and envision directions for collaborative work and contributions (Zhang et al., 2009). In parallel with such challenges for students, the teacher needs to actively follow the online discourse to understand the landscape of collective ideas, identify and assess advances in focal areas, and foster further efforts to investigate emerging and deeper issues. However, manual implementation of such analyses of online discourse is often labor-intensive and demanding even for the researchers, let alone the teachers or the students. This calls for new assessment and analysis tools to help students and their teacher trace online discourse over time and provide feedback on collective progress as well as individual participation. The purpose of this research is to test automated analysis based on machine learning techniques to discover and trace major topics of inquiry based on online discourse data. Such automated analysis may provide learners and teachers with ongoing assessment and feedback about the evolving landscape and status of their collective understanding achieved through online discourse; it also provides researchers with new and less labor-intensive tools to analyze online discourse.

Tracing and Analyzing Online Knowledge-Building Discourse as Unfolding Conceptual Streams

Diverse measures have been developed to trace and examine knowledge advancement in online discourse (Hmelo-Silver et al., 2007; van Aalst & Chan, 2007; Weinberger & Fischer, 2006), with efforts further made to automate some of the analyses (Law et al., 2007; Rosé et al., 2008; Teplovs & Fujita, 2009). To trace collective idea advancement in online discourse, Zhang (2004) developed inquiry threads analysis in light of the discourse flow theory that explains the dynamics giving directions to the flow of thoughts in discourse (Chafe, 1997). In interactive conversations, members dynamically maintain a shared focus of active consciousness, which constantly shifts and develops as the conversation proceeds. Interrelated focuses of conversation constitute larger focus clusters—discourse topics. It is often presented in the form of a problem or gap yet to be filled that engages members' thinking in an active and determinate psychic way. The evolving streams of the whole discourse are sustained through the joint interplay of the participants, whose successive focuses of consciousness propelled forward by their reaction to what others have just said (Chafe, 1997). The whole course of conversation unfolds in an emergent and improvisational manner. Inquiry threads analysis traces collective idea progress in extended discourse by identifying focal topics or thematic problems that have emerged. Discourse contributions that address each shared problem/theme (e.g. how are rainbows created) form into an inquiry thread, a conceptual stream of discourse extending from the first to the last discourse entry (Zhang et al., 2007). Advances in each inquiry thread are further elaborated through analyzing the content of major discourse input (e.g., questions, ideas). The scope and depth of the discourse is additionally benchmarked by comparing the themes of the inquiry threads against the curriculum expectations (Sun et al., 2010). A Web-based program—Idea Thread Mapper (ITM)—has been developed to help researchers create inquiry threads based on online discourse data as a means to examine collective knowledge progress. ITM has further been turned into a classroom tool to help teachers trace and assess collective progress in online discourse beyond individual postings and engage students in reflective monitoring of collective knowledge progress. Students and their teacher work together to identify focal knowledge themes and goals and to cluster discourse entries related to each theme as a conceptual thread of inquiry. Using the ITM tool, students can visualize a set of interrelated inquiry threads on a timeline, reflect on advances of knowledge over time, and identify weak areas as the focus of their further work (Chen et al., 2013). The current study explores automated analysis to identify major themes/topics of discourse as well as discourse entries related to each topic. We envision that such tools can facilitate students, teachers, and researchers to analyze and assess conceptual threads of inquiry unfolding in online discourse. Furthermore, they may also help in developing assessment of online discussions in general online educational contexts. Instead of simply evaluating forum posts as stand-along essays, instructors can use such analysis to obtain the conceptual landscape of the entire discussion (Dringus & Ellis, 2005) and further assess whole class progress as well as individual participation and contribution.

Existing Work to Analyze Online Discourse Using Probabilistic Topic Models

In educational data mining, automated content analysis has been tested as a major method for online discourse analysis (Rose´ et al., 2008; Mu et al., 2012; Baker & Yacef, 2009; Romero & Ventura, 2007; Romero & Ventura, 2010). Content analysis requires the design of pre-defined coding schemes focusing on various patterns of discourse of interest to the researcher. Drawing on some of the well-tested coding schemes focusing on patterns of argumentative discourse (e.g., Weinberger & Fischer, 2006), researchers have designed automated content analysis tools using nature language processing and machine learning techniques (Mu et al., 2012). The existing automated content analysis tools primarily focus on patterns of discourse at a single post level. They usually lack the ability to address conceptual themes of ideas that evolve in this whole course of the discourse.

As an alternative to content analysis, topic models that automate the discovery of thematic topics from a corpus have gained increasing attention. One appealing property of analysis based on topic models is that they are unsupervised algorithms, which obviates the requirement of manually annotating the corpus, and may help to reduce cost, and improve objectivity of the analysis results. The objective of probabilistic topic modeling is to automatically discover the topics in a corpus and the topic assignments of each document using only the documents in the corpus (more precisely, all the words, or the frequency of unique words). In the language of statistical analysis, we treat the documents as the observed variables, while the topic structures — the individual topics, the per-document topic distributions, and the per-document per-word topic assignments—are regarded as the latent variables in the model. The central computational problem in probabilistic topic modeling is to use the observed documents to infer the latent topic structure. This can be thought of as "reverse engineering" the simple generative process underlying the topic models — what is the latent topic structure that is likely to generate the observed collection. Two common choices of topic models are the latent semantic indexing (LSI) (Hofmann, 2001) and latent Dirichlet allocation (LDA) (Blei, 2012). In LDA, we assume that the conditional distribution of the topic assignment given the per-document topic proportion, the conditional distribution of the observed word given all the topics and the per-word topic assignment are multinomial distributions, while the prior distributions over the individual topics and per-document topic assignments are Dirichlet distributions. According to the Bayesian framework, the generative model reduces to compute the conditional distribution of the topics and topic assignments of each word and document given the observed corpus. In practice, we use approximation methods to evaluation of the document posterior distribution, the two main categories of which are variational methods and sampling based methods. Though both methods have been shown leading to reliable inference performances, in this work, we employ the variation-based method for its running efficiency. Compared with LDA, LSI has a significant drawback in that the "topics" recovered from LSI lacks a clear semantic meaning. On the other hand, the topics in LDA are distributions of words, and correspond to components in the probabilistic generative model. More importantly, LDA usually outperforms LSI in practice (Blei, 2012) mainly due to its better handling of synonymy and polysemy because of the probabilistic association of words to topic. Therefore in this study, we tested using LDA to discover and trace major topics of inquiry based on online discourse.

Early adoptions of topic models for educational data include the work of Ming and colleagues (2012), which applied two topic models, namely probabilistic LSI and hierarchical LDA, to predict the grades of the students and showed that these analyses provide information that aids more precise student assessment. Y. Zhang and colleagues (2012) applied LDA to online discussions of four Chinese classrooms to extract topics and display the temporal profiles of the topics. This study suggests that frames built from the top terms of the learned topics support easier human interpretation. Beyond online learning, Sherin (2012, in press) tested using LDA and latent semantic analysis to extract fragments (categories) of ideas from student interviews in order to code misconceptions versus scientific explanations. The results of the automated analysis aligned closely with the coding of human analysts. The above studies point to the promising potential of LDA to capture conceptual topics and structures in student discourse data. However, this potential needs to be further explored by applying the topic model analysis to organize online discourse of productive knowledge building communities that engage in emergent, progressive inquiry over a longer term to capture unfolding directions of collective knowledge work. Such automated analysis needs to be benchmarked by comparing the results with those from manual efforts of human analysts. Therefore, this exploratory study intends to use topic model analysis to examine unfolding processes of collective knowledge building in the online discourse of a Grade 4 knowledge building community and compare the results with human coding.

The Corpus of Online Discourse and Classroom Context

This research analyzed the online discourse of a class of 22 fourth-graders (9-to-10-year-olds) who studied light over a three-month period supported by Knowledge Forum a collaborative online knowledge building environment (Scardamalia & Bereiter, 2006). The students were taught by a teacher who had strong expertise in facilitating knowledge building, as indicated through a prior study that analyzed his improvement of knowledge building practices over three years (Zhang et al., 2009). The classroom design encouraged students to take on collective responsibility for high-order decision-making related to knowledge goals, long-range planning, and progress evaluation. Students generated problems of understanding, discussed diverse ideas and theories through face-to-face knowledge-building discourse, conducted self-generated experiments and observations, and searched libraries and the Internet. They contributed problems, ideas, data, and resources into Knowledge Forum for continual discourse and idea improvement. Knowledge Forum provided the public knowledge space in which student ideas and inquiry work were recorded, in views (workspaces) corresponding to their focal goals. By writing notes (discourse entries) in these views, the students contributed their ideas, data, and related information. Supportive features for knowledge-building discourse allowed the students to co-author, build on, and annotate notes; create reference links with citations to existing notes; and create rise-above notes to summarize, distil, and advance their discussions. Content analysis of student portfolio notes that summarized their knowledge advances showed significant progress made over the three months in understanding deep issues about light (see Zhang et al., 2010).

The corpus of online discourse contains 149 notes over a vocabulary of 824 distinct words, among which 75 words are stop words, namely, words that only assume grammatical functions or carry little meanings relevant to the analysis, such as articles, prepositions, and pronouns. After removal of the stop words, the number of meaningful distinct words is reduced to 749, with each note in the corpus containing 43 distinct words on average. Hereinafter, we use the terms 'note' and 'document' interchangeably depending on the context: we refer to discourse entries as notes in the context of Knowledge Forum, and as documents in the context of topic modeling.

Topics in online discourse can be defined at different levels of specificity, ranging from broader themes of discussion to specific topics and sub-topics. Different units of topic clustering may serve different needs: reflection of students and teachers on discourse progress tend to focus on relatively larger themes and topics while researchers often prefer more fine-grained analysis of topics and ideas. In the inquiry of light, the students and their teacher co-identified eight themes to guide their inquiry and discourse about light: Why is snow white? (reflection and absorption) How do plants adapt to light? Why is night cooler than daytime? How are shadows made? How does light travel? How are rainbows created? (colors of light) How do lenses work? And how do mirrors work? Elaborating on these eight themes, two researchers collaboratively identified 17 specific topics from student online discourse, and used this list of topics to code each Knowledge Forum note. Notes addressing the same topical issue constituted a conceptual line of inquiry--an inquiry thread—that extended from the first to the last document created (see Chen, 2013 for further details of this analysis). The result was the identification of 17 inquiry threads that map onto the eight overarching themes identified by the knowledge building community.

Topics Discovered with LDA Analysis

The eight overarching themes and 17 specific inquiry threads provided the contextual information needed for us to test topic discovery with LAD analysis. We tested a range of total number of topics to be discovered ranging from 5 to 17 topics, and setting the number to 10 topics generated the most interpretable result.

Table 1 shows the 10 topics discovered through the LDA analysis, with a list of top 10 keywords for each topic. The 'Keywords' column lists the vocabulary that has the largest β value under a certain topic, that is, the words that are mostly likely to belong to that topic. In the 'Interpretation' column, we present a summarization of each topic obtained by analyzing the keywords used in the documents that the algorithm assigned to the topic. Some of the topics (e.g. Topic 9) are harder to interpret than others. There are substantial overlaps (shared keywords) between topics 1 (Light travels through materials), 5 (Reflection) and 9 (Materials that reflect); and between topics 3 (Shadows, including colored shadows) and 8 (Shadows and light sources).

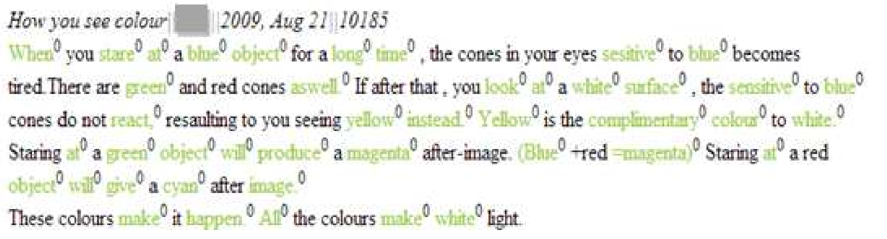
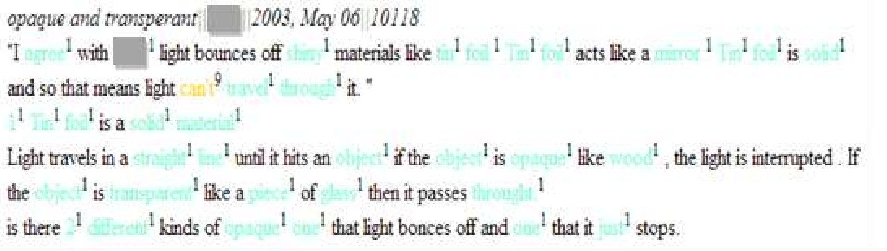
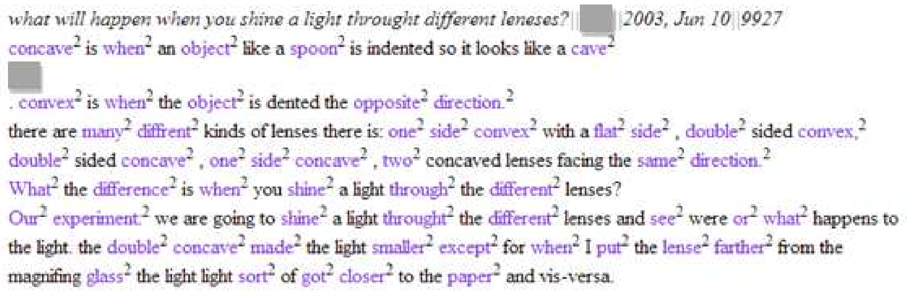
As we navigated through the results from our test with M = 5, 6…17 topics, we found that some topics are interpretable at certain Ms but lost their interpretability as the parameter increases or decreases.

Table 1: Ten Topics Extracted by LDA, Each with the Top Keywords and an Interpretation.

Topic	Keywords	Interpretation
Topic 0	'colour' 'r' 'green' 'yellow' 'make' 'blue' 'object' 'cone' 'primary' 'at'	Colors of light
Topic 1	'tin' 'foil' 'solid' 'glass' 'travel' 'through' 'material' 'solstice' 'can' 'mean'	Light travels (through materials)
Topic 2	'mirror' 'convex' 'when' 'concave' 'reflection' 'side' 'lens' 'telescope' 'two' 'see'	Mirrors and lenses
Topic 3	'rainbow' 'when' 'shadow' 'color' 'made' 'glass' 'through' 'colour' 'can' 'think'	Shadows (including colored shadows)
Topic 4	'glass' 'what' 'see' 'eye' 'solid' 'when' 'people' 'through' 'very' 'back'	See
Topic 5	'mirror' 'shine' 'reflect' 'direction' 'will' 'line' 'plant' 'this' 'work' 'bounce'	Mirrors and reflection
Topic 6	'sun' 'when' 'earth' 'moon' 'eclipse' 'shadow' 'other' 'world' 'around' 'line'	Eclipses and seasons
Topic 7	'white' 'snow' 'colour' 'prism' 'black' 'melt' 'when' 'see' 'fast' 'why'	Snow and white light
Topic 8	'shadow' 'object' 'made' 'opaque' 'energy' 'part' 'call' 'umbra' 'what' 'go'	Shadows and light sources
Topic 9	'through' 'go' 'can' 'reflect' 'tinfoil' 't' 'think' 'was' 'angle' 'when'	Materials

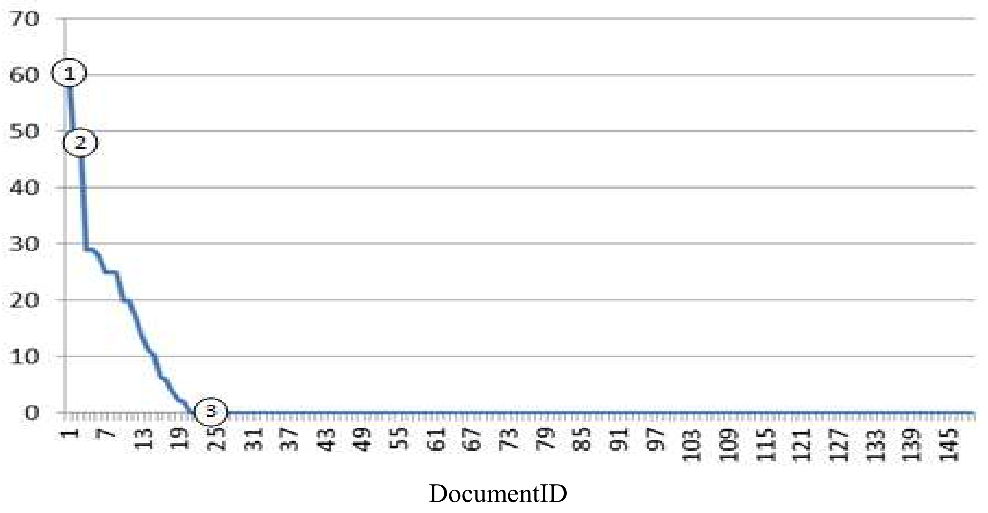
Table 2 displays some example documents for the first three topics. Aligned with the interpretation, these documents discuss colors, light traveling through materials, and mirrors and reflection, respectively. The documents in Table 2 are structured as the following: the first line of the documents lists the title, author initials and document creation date information in italic font separated by '||'; the contents of the documents are shown in the remaining lines. The different font color and superscripts represent the topic assignment of each word. For example, a word in green font with superscript 0 means that the topic assignment of this word is Topic 0. We refer the readers to the online supplementary tables for more examples.

Table 2. Example Documents for the First Three Topics

Topic 0 Colors of light	 *How you see colour|* [redacted] *||2009, Aug 21||10185* When^{0} you stare^{0} at^{0} a blue^{0} object^{0} for a long^{0} time^{0} , the cones in your eyes sesitive^{0} to blue^{0} becomes tired.There are green^{0} and red cones aswell.^{0} If after that , you look^{0} at^{0} a white^{0} surface^{0} , the sensitive^{0} to blue^{0} cones do not react,^{0} resaulting to you seeing yellow^{0} instead.^{0} Yellow^{0} is the complimentary^{0} colour^{0} to white.^{0} Staring at^{0} a green^{0} object^{0} will^{0} produce^{0} a magenta^{0} after-image. (Blue^{0} +red =magenta)^{0} Staring at^{0} a red object^{0} will^{0} give^{0} a cyan^{0} after image.^{0} These colours make^{0} it happen.^{0} All^{0} the colours make^{0} white^{0} light.
Topic 1 Light travels	 *opaque and transperant|* [redacted] *|2003, May 06||10118* "I agree^{1} with [redacted]^{1} light bounces off shiny^{1} materials like tin^{1} foil ^{1} Tin^{1} foil^{1} acts like a mirror ^{1} Tin^{1} foil^{1} is solid^{1} and so that means light can't^{9} travel^{1} through^{1} it. " 1^{1} Tin^{1} foil^{1} is a solid^{1} material^{1} Light travels in a straight^{1} line^{1} until it hits an object^{1} if the object^{1} is opaque^{1} like wood^{1} , the light is interrupted . If the object^{1} is transparent^{1} like a piece^{1} of glass^{1} then it passes throught.^{1} is there 2^{1} different^{1} kinds of opaque^{1} one^{1} that light bonces off and one^{1} that it just^{1} stops.
Topic 2 Mirrors and lenses	 *what will happen when you shine a light throught different leneses?||* [redacted] *|2003, Jun 10||9927* concave^{2} is when^{2} an object^{2} like a spoon^{2} is indented so it looks like a cave^{2} [redacted] . convex^{2} is when^{2} the object^{2} is dented the opposite^{2} direction.^{2} there are many^{2} diffrent^{2} kinds of lenses there is: one^{2} side^{2} convex^{2} with a flat^{2} side^{2} , double^{2} sided convex,^{2} double^{2} sided concave^{2} , one^{2} side^{2} concave^{2} , two^{2} concaved lenses facing the same^{2} direction.^{2} What^{2} the difference^{2} is when^{2} you shine^{2} a light through^{2} the different^{2} lenses? Our^{2} experiment.^{2} we are going to shine^{2} a light throught^{2} the different^{2} lenses and see^{2} were or^{2} what^{2} happens to the light. the double^{2} concave^{2} made^{2} the light smaller^{2} except^{2} for when^{2} I put^{2} the lense^{2} farther^{2} from the magnifing glass^{2} the light light sort^{2} of got^{2} closer^{2} to the paper^{2} and vis-versa.

LDA-Based Analysis of Topic-Document Relationships

The LDA algorithm further outputs the assignment of topics to documents. Figure 2 shows how likely each of the 149 documents belongs to Topic 0 that was interpreted as colors of light. The x-axis lists all the 149 documents, and the y-axis shows the score of how much each document pertains to the topic 'color', with higher numerical score indicating greater relevance. We show the content of three documents of different likelihood scores in Figure 2. The first two documents have higher likelihood scores, and their contents are more pertaining to colors, while the last document has lower likelihood score and its content is not related to colors.



○ *How you see colour||X.X.||2009, Aug 21||10185*

When you stare at a blue object for a long time, the cones in your eyes sesitive to blue becomes tired. There are green and red cones as well. If after that, you look at a white surface, the sensitive to blue cones do not react, resaulting to you seeing yellow instead. Yellow is the complimentary colour to white. Staring at a green object will produce a magenta after-image. (Blue +red =magenta) Staring at a red object will give a cyan after image. These colours make it happen. All the colours make white light.

○ *How do we see objects that are coloured||X.X. , X.X.||2003, Jun 02||10356*

A green object looks green because light hits it and all the green of the light gets reflected into you'r eye but all the other colours get absorbed into the object. The cones in your eyes tell your brain what the colour the object is. If the colour is not primary the light hits the cones in your eyes which mixes the colours that make up the single colour (E.G. yellow is made by red and green) which apears as the single colour.

○ *But Glass is solid too!||X.X.||2003, May 05||10112*

"I agree with XX light bounces off shiny materials like tin foil. Tin foil acts like a mirror. Tin foil is solid and so that means light can't travel through it. " Tin foil is a solid material I disagree with your theory XX because glass is a solid material too, and light can go through it!

Figure 2: Assignment of Topic 0 over all documents. This diagram plots how likely each document belongs to Topic 0. The circled numbers, ①, ② and ③mark the position of the likelihood of the three documents shown under the graph.

Comparing Document-Topic Assignment with LDA against Human Coding

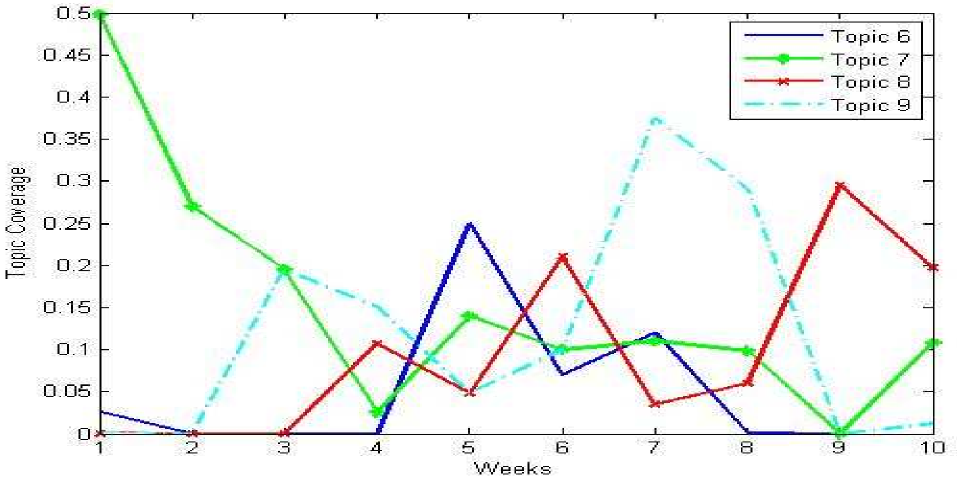
The results reported above show how the LDA algorithm can extract topics from online discourse and assign interpretable topics to each discourse entry. To further gauge the accuracies of these topic assignments, we compared the LDA assignments with those obtained with manual coding.

The evaluation process was as follows: we randomly selected five of the ten topics and pool the top six documents from each topic. The order of the documents was then randomized. Two human raters (two of the authors) independently read each of the thirty documents and rated the relevance of each documents to the five topics using a 7-point Likert scale (from 0-definitely not related to 6-definitely related). We then compared the algorithm's topic assignments against the average of the human raters' results.

We used two evaluation metrics: Fleiss Kappa (Fleiss, Levin & Paik, 2013) and normalized Discounted Cumulative Gain (nDCG) (Järvelin & Kekäläinen, 2002). The Fleiss Kappa score describes how much the rater/algorithm's answers differ from random assignment of topics. Hence, the Fleiss kappa tells us the percentage of agreement gained beyond chance. We used it to evaluate the agreements on the assignment of the most relevant topics to the documents. Kapper for inter-rater agreement is 1, and for system-human consistency is 0.87. The nDCG measure compares the similarity between two lists, range from 0 to 1. The nDCG evaluation metric is suitable for our testing as it captures both the importance of the ranking and the numerical gains that our human annotators assigned. Considering that our system outputs at most 2 topics for each document (in fact, for only 4 documents, our system output 2 topics, and it assigned only 1 topic for the rest of the documents), we only calculated the result for the selecting the most relevant 1 topic, and the most relevant 2 topics. For the most relevant topic, nDCG (averaged over all 30 documents) for inter-rater agreement is 1, and for system-human consistency is 0.90. For the two most relevant topics, the inter-rater agreement in terms of nCDG (averaged over all 30 documents) is 0.99, and the system-human consistency is 0.86. These results indicate that the topic assignment generated by the LDA algorithm achieved an acceptable agreement with human judgment, even though the agreement is lower than that between the two human coders.

Analyzing Temporal Evolution of Different Topics in the Online Discourse

We further tested how LDA may be used to generate useful analysis and feedback data for educators and researchers by examining the progressive changes in student online discourse. Figure 3 shows the evolution of four topics over the 10-week period of inquiry. The x-axis represents time in term of weeks (week 1 – 10), and the y-axis shows how prominent the topic is in that week's discussion (accumulated γ scores for all the posts within given week). For the sake of clarity, we only plotted the scores for four topics in Figure 3.



Topic 6	'sun' 'when' 'earth' 'moon' 'eclipse' 'shadow' 'other' 'world' 'around' 'line' Eclipses and seasons	Eclipses and seasons
Topic 7	'white' 'snow' 'colour' 'prism' 'black' 'melt' 'when' 'see' 'fast' 'why'	White light and snow
Topic 8	'shadow' 'object' 'made' 'opaque' 'energy' 'part' 'call' 'umbra' 'what' 'go'Shadows and light sources	Shadows and light sources
Topic 9	'through' 'go' 'can' 'reflect' 'tinfoil' 't' 'think' 'was' 'angle' 'when'	Reflective materials

Figure 3. Temporal Projections of Topics.

The temporal progress of the topics indicates many interesting aspects of the learning process. For instance, Topic 7 (snow and white light) has a dominant score during the first week, and decreases over the next few weeks, then rises again in week 5. The intensive discourse about this topic in the first week as detected by LDA coincides with what actually happened in the classroom: at the beginning of the light inquiry, an early spring snow triggered students' interest in why snow is white and what would happen if it were black. These issues became the primary focus in the first week in the online and face-to-face activities, with 9 notes written in the first week (see the inquiry thread Snow in Figure 1). These issues related to snow became less central in the following three weeks as the knowledge building community formulated other, deeper themes of inquiry to address a wide range of optical issues. These deeper issues were represented by the increase of Topics 8 and 9 in

Figure 3 as well as the inquiry threads on primary and secondary colors, light travels and interacts with different materials, mirrors and reflection, shadows, and how we see things in Figure 1. Interestingly, the inquiry on primary and secondary colors and how we see colors led students to understanding the nature of white light and how it can be split into different colors using a prism, with a number of new notes written in week 5-10, as detected by the LDA analysis.

Discussion

This study of applying LDA to the online discourse data showed promising results. The algorithm was able to detect semantically meaningful topics even when the target corpus has a relatively limited vocabulary and short notes and addresses coherently connected issues in one content area (optics) instead of different areas (e.g. optics vs. chemistry). The ten topics discovered by LDA reflect core optical issues investigated by the knowledge building community. These topics overlap substantially with —although do not map directly onto— the inquiry themes identified by students and inquiry thread topics identified by researchers.

An important aspect in applying LDA is to determine the possible number of topics before running LDA. In the current study, we identified the appropriate number of topics by experimenting with varying number of topics based on a possible range informed by the reflection of students on their collective knowledge themes as well as the manual coding of researchers. The researchers in this study implemented detailed coding of notes using inquiry threads to understand the process of knowledge building and benchmark automated analysis results. However, in future application of LDA analysis, such manual review can be handled more quickly to identify major topics of discussion without document-by-document coding. With an overall sense of the possible topics discussed, the analyst—researcher or educator--can vary the total number of topics to used in the LDA algorithm and interpret the obtained candidate topics by examining the keywords involved, with keywords shared between different topics showing the conceptual connections. It is important to point out that most of the topics detected in this study are hard to interpret accurately based on stand-alone keywords without looking at how these words are used in the context of the corresponding discourse entries. Future design of LDA-base assessment tools should show the keywords of each topic in context (e.g. sentences) to aid user interpretation.

Using the LDA algorithm, we further identified documents that are likely associated with each of the ten topics detected. Comparing the document-topic assignment by LDA against the judgment of two human coders resulted in high-level consistency measured using nCDG and Kappa. Such automated document-topic assignment may help students and teachers to identify important discourse contributions addressing particular focal themes as a way to review collective progress and individual contribution. It may also aid researchers in coding discourse entries based on conceptual topics to understand the evolving scope and depth of inquiry evident in the online discourse. Meanwhile, it is important to mention that our evaluation of automated document-topic assignment was based on a small sample of most relevant documents for five topics, with a high level of accuracy. The accuracy may decrease when assigning other documents that have lower relevance (i.e. likelihood) scores; in such cases, LDA may provide a list of candidate documents that are potentially relevant to each topic for the human analyst to choose from, so the analyst can select relevant documents for each topic more easily.

We also noticed a few challenges, which suggest headroom for further improvement of the technology. The unsupervised-learning nature of the LDA algorithm provides researchers little control over the granularity and structure of the detected topics. The algorithm may generate un-interpretable topics and it is difficult to estimate the interpretability of the generated topics without human interference. Sometimes the educator or analyst may want to see a finer-grain analysis of certain topics but it is challenging to control the granularity of the topics detected by LDA. To address these drawbacks, future research may look into developing and applying topic models that can incorporate information from manually-coded archived online discourse on the same topic. Also, we will also explore the use of hierarchical topic models that can discover topics with intrinsic hierarchical orders to solve the granularity issue. We believe this further development will give the educator and analyst more control over the discovered topics for the purpose of more effectively and efficiently analyzing large volumes of online discourse data, and pave the road for creating auto-assessment and feedback tools to leverage sustained and productive knowledge-building discourse.

References

Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*(1), 3-17.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Chen, M.-H., Zhang, J. & Lee, J. (2013). Making Collective Progress Visible for Sustained Knowledge Building. In N. Rummel, M., Kapur, M. Nathan, & S. Puntambekar (Eds.), To See the World and a Grain of Sand: Learning across Levels of Space, Time, and Scale: CSCL 2013 Conference Proceedings Volume 1 (pp.81-88). International Society of the Learning Sciences.

Chafe, W. (1997). Polyphonic topic development. In T. Givón (Ed.), *Conversation: Cognitive, Communicative and Social Perspectives* (pp. 41-53). John Benjamins Publishing.

Dringus, L. P. and Ellis, T. (2005). Using data mining as a strategy for assessing asynchronous discussion forums. *Computers & Education*, 45(1):141–160.

Fleiss, J. L., Levin, B., & Paik, M. C. (2013). *Statistical methods for rates and proportions*. Hoboken, NJ: John Wiley & Sons.

Hmelo-Silver, C.E. (2003). Analyzing collaborative knowledge construction: Multiple methods for integrated understanding. *Computers & Education*, 41, 397-420.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, *42*(1-2), 177-196.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, *20*(4), 422-446.

Law, N., J. Yuen, Huang, R., Li, Y. & Pan, N. (2007). A Learnable Content and Participation Analysis Toolkit for Assessing CSCL Learning Outcomes and Processes. In C. A. Chinn, G. Erkens, & S. Puntambekar (Eds.), *Proceedings of the 8th International Conference on Computer Supported Collaborative Learning* (pp. 411-420). International Society of the Learning Sciences.

Meier, A., Spada, H., Rummel, N. (2007). A rating scheme for assessing the quality of computer-supported collaboration process. *International Journal of Computer-Supported Collaborative Learning*, 2, 63-86.

Ming, N., & Ming, V. (2012). Predicting student outcomes from unstructured data. *Proceedings of the 20th Conference on User Modeling, Adaptation, and Personalization.* Montreal, Canada

Mu, J. Stegmann, K., Mayfield, E., Rose, C., & Fischer, F. (2012). The ACODEA framework: Developing segmentation and classification schemes for fully automated analysis of online discussions. *International Journal of Computer-Supported Collaborative Learning*, 7, 285-305.

Rosé, C., Wang, Y.-C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., Fischer, F. (2008). Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in CSCL. *International Journal of Computer-Supported Collaborative Learning*, 3(3), 237-271.

Scardamalia, M., & Bereiter, C. (2006). Knowledge building: Theory, pedagogy, and technology. In K. Sawyer (Eds.), *Cambridge Handbook of the Learning Sciences* (pp. 97-118). New York, NY: Cambridge University Press.

Sherin, B. (2012). Computing student science conceptions with Latent Dirichlet Allocation. *Paper presented at the 10th International Conference of the Learning Sciences*, Sydney, Australia.

Sherin, B. (in press). A Computational Study of Commonsense Science: An Exploration in the Automated Analysis of Clinical Interview Data. *Journal of the Learning Sciences.*

Stahl, G. (2006). Group cognition: Computer support for building collaborative knowledge. Cambridge, MA: MIT Press.

Sun, Y., Zhang, J., & Scardamalia, M. (2010). Knowledge building and vocabulary growth over two years, Grades 3 and 4. Instructional Science, 38(2), 247-271.

Teplovs, C., & Fujita, N. (2009). Determining curricular coverage of student contributions to an online discourse environment through the use of latent semantic analysis and term clouds. *Proceedings of the 9th international conference on Computer supported collaborative learning* (Vol. 2, pp. 165-167). International Society of the Learning Sciences.

van Aalst, J., & Chan, C.K..K. (2007). Student-directed assessment of knowledge building using electronic portfolios in Knowledge Forum. *Journal of the Learning Sciences*, 16, 175-220.

Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education*, 46, 71-95.

Zhang, J. (2004). The growing networks of inquiry threads in a knowledge building environment. Paper presented at the Knowledge Building Summer Institute. Ontario Institute for Studies in Education, University of Toronto.

Zhang, J., & Messina, R. (2010). Collaborative productivity as self-sustaining processes in a Grade 4 knowledge building community. In K. Gomez, J. Radinsky, & L. Lyons (Eds.), Proceedings of the 9th International Conference of the Learning Sciences (pp. 49-56). Chicago, IL: International Society of the Learning Sciences.

Zhang, J., Scardamalia, M., Lamon, M., Messina, R., & Reeve, R. (2007). Socio-cognitive dynamics of knowledge building in the work of nine- and ten-year-olds. Educational Technology Research and Development, 55(2), 117–145.

Zhang, J., Scardamalia, M., Reeve, R., & Messina, R. (2009). Designs for collective cognitive responsibility in knowledge building communities. Journal of the Learning Sciences, 18(1), 7–44.

Zhang, Y., Law, N., Li, Y., and Huang, R. (2012). Automatic extraction of interpretable topics from online discourse. In *The International Conference of the Learning Sciences (ICLS) 2012, Volume 1*, (p.443-450). International Society of the Learning Sciences.