Analyzing Online Learning Discourse using Probabilistic Topic Models

Weiyi Sun
Department of Informatics
University at Albany, SUNY
Albany, NY 12222
wsun2@albany.edu

Siwei Lyu
Department of Computer Science
University at Albany, SUNY
Albany, NY 12222
lsw@cs.albany.edu

Hui Jin
School of Education
University at Albany, SUNY
Albany, NY 12222

Jianwei Zhang
School of Education
University at Albany, SUNY
Albany, NY 12222
jzhang1@albany.edu

Abstract

This exploratory study applied probabilistic topic models to analyze the online discourse over the topic of optics among a group of Grade 4 students. Using the Latent Dirchilet Allocation (LDA) model, we extract ten distinct and semantically meaningful clusters (i.e., topics) from the online discourse, which overlap substantially with —although do not map directly onto—the inquiry themes identified by students and researchers. The LDA analysis further identifies 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. Further analysis with LDA helps to trace the evolution of different topics over time and compare student discourse against the expectations of the curriculum. These results suggest the potential of LDA to help trace and assess online discussions in collaborative learning settings and online courses.

1 Introduction

With online learning increasingly adopted across all levels of education, researchers and practitioners seek effective ways to make wise use of the plethora of online data to trace and leverage student learning. Supported by collaborative online environments, such as Knowledge Forum (Scardamalia & Bereiter, 2006), students engage in semester-long asynchronous discourse to contribute and refine ideas, address deepening questions, and advance their collective understanding. Meanwhile, the teachers need to actively follow the online discourse to understand the 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. 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.

Drawing on existing efforts to manually analyze conceptual advances in online discourse (Zhang et al., 2007), this research further tests automated analysis based on probabilistic topic models 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 of their collective understanding achieved through online discourse; it also provides researchers with new and automated tools to analyze discourse in online education settings.

2. Previous work

Applying data mining techniques to educational data becomes a popular research topic in the field of the learning sciences (Rose´ et al., 2008; Mu et al., 2012; Baker & Yacef, 2009; Romero & Ventura, 2007; Romero & Ventura, 2010). Topic models, such as Latent semantic indexing (LSI) (Hofmann, 2001) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003), due to their unsupervised learning natures, have gained increasing attention in the research community of educational data mining and machine learning. Early adoptions of topic models for educational data include the work of Ming et al. (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 mentioned 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 validated by online discourse of productive knowledge building communities to capture unfolding directions of collective knowledge work. We also need to benchmark it against manual coding of human analysts. Therefore, this 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. Our preliminary results suggest wider applicability of topic models in educational data mining, whenever the task predicates on the extraction or assignment of high-level thematic topics.

3. Method

3.1 Latent Dirichlet Allocation (LDA)

Assuming a corpus with D documents, each containing N words[1] to be represented with K topics, which we denote as $b_{1:K}$ with each being a distribution over the vocabulary. The topic proportions for the dth document are c_d, where $c_{d,k}$ is the topic proportion for topic k in document d. The topic assignments for the dth document are z_d, where $z_{d,n}$ is the topic assignment for the nth word in document d. Last, the observed words for document d are denoted as a vector w_d, where $w_{d,n}$ is the nth word in document d, which is an element from the fixed vocabulary. With these notations, the generative model of LDA, as described previously, corresponds to the following joint probability distribution over the latent and observed variables:

$$p(b_{1:K}, c_{1:D}, z_{1:D}, w_{1:D}) = \prod_{k=1}^{K} p(b_k) \prod_{d=1}^{K} p(c_d) \prod_{n=1}^{N} p(z_{d,n}|c_d) p(w_{d,n}|b_{1:K}, z_{d,n})$$

This joint probability distribution is fully specified in LDA(Blei et al., 2003), where the conditional distribution of the topic assignment $z_{d,n}$ given the per-document topic proportion c_d and the conditional distribution of the observed word given all the topics $b_{1:K}$ and the per-word topic assignment $z_{d,n}$ are multinomial distributions, while the prior distributions over the individual topics b_k and per-document topic assignments c_d are Dirichlet distributions. According to the Bayesian framework, this reduces to compute the conditional distribution of the topics and topic

[1]We assume here for simplicity that all documents have the same number of words, but it is not difficult to handle the general case when each document may have different number of words.

assignments of each word and document given the observed corpus. In practice, precise evaluation of the document posterior distribution is intractable. Hence, we resort to approximation methods to tackle this problem, 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.

The purpose of this study is to test using LDA to discover thematic topics emerged from extended online knowledge-building discourse, identify major discourse entries addressing each topic, and analyze discourse contributions and advances over time. Therefore, the specific approach tested through this study serves to achieve four interconnected goals: to organize large corpus of online discourse by topics, to retrieve relevant discourse entries by matching topic assignments, to conduct temporal analysis of topic evolution, and to compare the discourse of students against the curriculum expectations.

3.2 Data Source 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 corpus contains 149 documents 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 document in the corpus containing 43 distinct words on average.

4. Results

Zhang & Messina (2010) conducted a manual analysis over the same corpus and identified eight overarching themes and 17 specific inquiry threads. Hence, we tested a range of total number of topics to be discovered ranging from 5 to 17 topics, and found that setting the number to 10 topics generated the most interpretable result.

The list of topics and keywords can be found in Table 1 of the Appendix. 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
Topic 2	'mirror' 'convex' 'when' 'concave' 'reflection' 'side' 'lens' 'telescope'	Mirrors and lenses
Topic 3	'rainbow' 'when' 'shadow' 'color' 'made' 'glass' 'through' 'colour' 'can'	Shadows /colored
Topic 4	'glass' 'what' 'see' 'eye' 'solid' 'when' 'people' 'through' 'very' 'back'	See
Topic 5	'mirror' 'shine' 'reflect' 'direction' 'will' 'line' 'plant' 'this' 'work'	Mirrors and reflection
Topic 6	'sun' 'when' 'earth' 'moon' 'eclipse' 'shadow' 'other' 'world' 'around'	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
Topic 9	'through' 'go' 'can' 'reflect' 'tinfoil' 't' 'think' 'was' 'angle' 'when'	Materials

Table 2 of the appendix 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.

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. " i^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.

5. Evaluation

To gauge the accuracies of these topic assignments, we compare the LDA assignments with those obtained with manual coding. The evaluation process is as follows: we selected five of the ten topics and pool the top six documents from each topic. The order of the documents is then randomized. Two human raters 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 compare the algorithm's topic assignments against the average of the human raters' results. We use two evaluation metrics: normalized Discounted Cumulative Gain (nDCG) (Järvelin & Kekäläinen, 2002) and Fleiss Kappa (Fleiss, Levin & Paik, 2013).

Considering that our system outputs at most 2 topics for each document, we only calculated the result for the selecting the most relevant 1 and 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. Kappa for inter-rater agreement is 1, and for system-human consistency is 0.87. The evaluation result shows 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.

6. Application of analysis results

6.1 Analyzing Temporal Evolution of Different Topics in the Online

Discourse

The analysis results may be used to generate useful analysis and feedback data for educators and researchers by examining the progressive changes in student online discourse. Figure 1 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 1.

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 both online and face-to-face activities and 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 results show the promising potential of LDA analysis to trace topic evolution in online discourse over time.

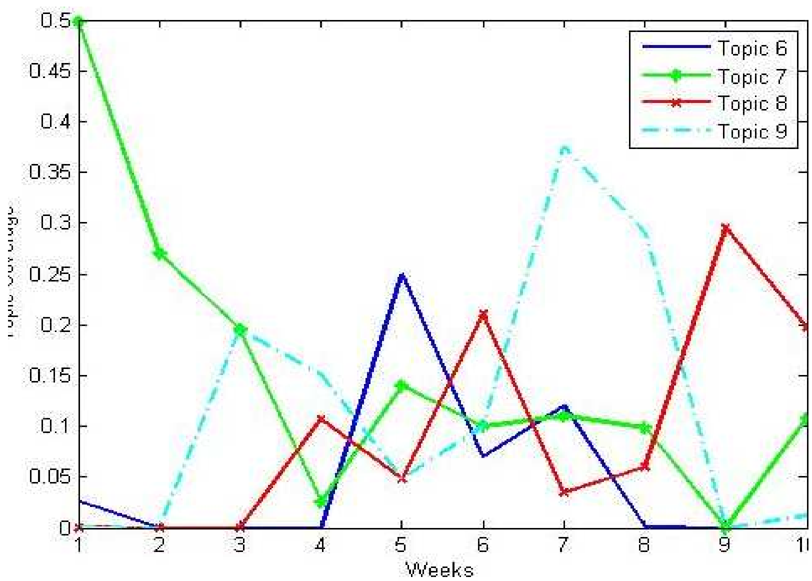

Figure 1 temporal projection of topics

6.2 Using LDA Results to Compare Student Discussion against Curriculum Guidelines

We may also utilize the analysis result to compare student discussions against the curriculum guidelines to identify strong as well as under-represented areas. This was achieved by applying the topic-word distribution computed by the LDA algorithm to the text of the Ontario Curriculum to estimate the coverage of the contents by the discussions. The Ontario Curriculum addresses light-related concepts first in Grade 4 (together with sound) and, then, more intensively in Grade 10. Figure 2 shows the estimated coverage of the curriculum for Grade 4 and 10 by student online discourse in Knowledge Forum. Consistent with our expectation, the analysis detected more overlap of the students' online discourse with the Grade 4 curriculum than with Grade 10 curriculum about optics.

	Requirements	topic 0	topic 1	topic 2	topic 3	topic 4	topic 5	topic 6	topic 7	topic 8	topic 9
1	investigate the basic properties of light (e.g., conduct experiments to show that light travels in a straight path, that light reflects off of shiny surfaces, that light refracts [bends] when passing from one medium to another, that white light is made up of many colours, that light diffracts [bends and spreads out] when passing through an opening)	-640.8	-582.9	-623.8	-604.4	-651.3	-619.1	-652.3	-597.1	-648	-601
2	use technological problem-solving skills (see page 16) to design, build, and test a device that makes use of the properties of light (e.g., a periscope, a kaleidoscope) or sound (e.g., a musical instrument, a sound amplification device) Sample guiding questions: How might you use what you know about sound or about light and mirrors in your device? Which properties of light or sound will be most useful to you in your device? What challenges might you encounter, and how can you overcome them?	-685.2	-693.2	-648.1	-679.6	-686.1	-678.9	-696.4	-681.6	-674.6	-674.6
3	use scientific inquiry/research skills (see page 15) to investigate applications of the properties of light or sound (e.g., careers where knowledge of the properties of light and/or sound play an important role [photography, audio engineering]; ways in which light and/or sound are used at home, at school, and in the community; ways in which animals use sound)	-636.8	-650.7	-624.8	-609.2	-622.9	-635.2	-667.4	-621	-641.6	-652.1
4	use appropriate science and technology vocabulary, including natural , artificial , beam of light , pitch , loudness , and vibration , in oral and written communication use a variety of forms (e.g., oral, written, graphic, multimedia) to communicate with different audiences and for a variety of purposes (e.g., create a song or short drama presentation for younger students that will alert them to the dangers of exposure to intense light and sound identify a variety of natural light sources (e.g., the sun, a firefly) and artificial light sources (e.g., a candle, fireworks, a light bulb)	-726.5	-722.4	-718.4	-709.8	-730.7	-730.8	-728.1	-719.5	-733.3	-721.7
5	distinguish between objects that emit their own light (e.g., stars, candles, light bulbs) and those that reflect light from other sources (e.g., the moon, safety reflectors, minerals)	-586.2	-668	-640.7	-691.5	-633.8	-679.1	-672.4	-694.9	-633.5	-638.3
6	describe properties of light, including the following: light travels in a straight path; light can be absorbed, reflected, and refracted	-677.8	-647.5	-581.7	-678.6	-566.1	-581.9	-679.3	-616.9	-668.7	-578.8
7	explain how vibrations cause sound describe how different objects and materials interact with light and sound energy (e.g., prisms separate light into colours; voices echo off mountains; some light penetrates through wax paper; sound travels further in water than air)	-661.5	-591.2	-611.4	-588	-604.9	-627.5	-719	-573.5	-600.6	-590.3
8	distinguish between sources of light that give off both light and heat (e.g., the sun, a candle, an incandescent light bulb) and those that give off light but little or no heat (e.g., an LED, a firefly, a compact fluorescent bulb, a glow stick)	-598.2	-677.9	-706.9	-642.1	-629.6	-705.3	-611.6	-605.4	-645.7	-630.4
9	identify devices that make use of the properties of light and sound (e.g., a telescope, a microscope, and a motion detector make use of the properties of light; a microphone, a hearing aid, and a telephone handset make use of the properties of sound follow established safety procedures for protecting eyes and ears (e.g., use proper eye and ear protection when working with tools)	-705.5	-743.9	-670.3	-689.6	-736.9	-686.9	-740.8	-702.3	-718.8	-728.5
10	investigate the basic properties of light (e.g., conduct experiments to show that light travels in a straight path, that light reflects off of shiny surfaces, that light refracts [bends] when passing from one medium to another, that white light is made up of many colours, that light diffracts [bends and spreads out] when passing through an olspening)	-640.8	-582.9	-623.8	-604.4	-651.3	-619.1	-652.3	-597.1	-648	-601
11	use technological problem-solving skills (see page 16) to design, build, and test a device that makes use of the properties of light (e.g., a periscope, a kaleidoscope) or sound (e.g., a musical instrument, a sound amplification device) Sample guiding questions: How might you use what you know about sound or about light and mirrors in your device? Which properties of light or sound will be most useful to you in your device? What challenges might you encounter, and how can you overcome them?	-685.2	-693.2	-648.1	-679.6	-686.1	-678.9	-696.4	-681.6	-674.6	-674.6
12	use scientific inquiry/research skills (see page 15) to investigate applications of the properties of light or sound (e.g., careers where knowledge of the properties of light and/or sound play an important role [photography, audio engineering]; ways in which light and/or sound are used at home, at school, and in the community; ways in which animals use sound)	-636.8	-650.7	-624.8	-609.2	-622.9	-635.2	-667.4	-621	-641.6	-652.1

Figure 2 – a Curriculum Coverage of Grade 4

	Requirements	topic 0	topic 1	topic 2	topic 3	topic 4	topic 5	topic 6	topic 7	topic 8	topic 9
13	analyse a technological device or procedure related to human perception of light (e.g., eye-glasses, contact lenses, infrared or low light vision sensors, laser surgery), and evaluate its effectiveness Sample issue: Laser surgery corrects vision by surgically reshaping the cornea to correct re-fractive defects in the eye. While the procedureis effective in most cases, it poses risks and can in some cases lead to poor night vision. Sample questions: How do anti-glare nightvision glasses help people who have difficulty driving at night? How do eyeglasses with colour filters help people with dyslexia to read?	-666	-669	-648.5	-663.1	-623.6	-682.1	-660.1	-649	-664.2	-648.1
14	analyse a technological device that uses the properties of light (e.g., microscope, retro-reflector, solar oven, camera), and explain how it has enhanced society [AI, C] Sample issue: Cameras can produce a range of optical effects, from highly detailed and realistic to manipulated and abstract. Photographic images are used for a wide range of purposes that benefit society, including in the areas of culture, education, security, policing, entertainment, and the environment. However, the wide spread use of cameras raises privacy concerns Sample questions: How do vision sensors help the Canadian Food Inspection Agency improve food safety? How are photonics used in the early diagnosis of diseases such as cancer? How have optical fibres enhanced our ability to communicate information? How do all of these technologies benefit society? How are outdoor lights such as street or stadium lights designed to limit light pollution in surrounding areas?	-743	-747.9	-724.8	-728.1	-744.3	-739.4	-739.6	-741.3	-739.9	-740.3
15	use appropriate terminology related to light and optics, including, but not limited to: angle of incidence, angle of reflection, angle of refraction, focal point, luminescence, magnification, mirage,and virtual image [C]	-717.1	-687.2	-618.7	-664.3	-668.1	-654.1	-727.7	-739	-693	-634.6
16	use an inquiry process to investigate the laws of reflection, using plane and curved mirrors, and draw ray diagrams to summarize their findings	-714.5	-607.2	-569.6	-636.1	-614.6	-614.6	-664.8	-610.3	-632.6	-573.9
17	predict the qualitative characteristics of images formed by plane and curved mirrors (e.g., location, relative distance, orientation, and size in plane mirrors; location, orientation, size, type in curved mirrors), test their predictions through inquiry, and summarize their findings [PR, AI, C]	-681.5	-632.4	-631	-697.4	-676.5	-667.1	-742.9	-660.1	-686.1	-633.6
18	use an inquiry process to investigate the refraction of light as it passes through media of different refractive indices, compile data on their findings, and analyse the data to determine if there is a trend (e.g., the amount by which the angle of refraction changes as the angle of incidence increases varies for media of different refractive indices) [PR, AI, C]	-742.7	-701.6	-690	-671.7	-716.6	-707.7	-716.8	-708.3	-725.7	-696.4
19	predict, using ray diagrams and algebraic equations, the position and characteristics of an image produced by a converging lens, and test their predictions through inquiry [PR, AI, C]	-725.7	-674.8	-647.7	-646	-669.8	-697.6	-691.1	-642.8	-694.1	-623.1
20	calculate, using the indices of refraction, the velocity of light as it passes through a variety of media, and explain the angles of refraction with reference to the variations in velocity [PR, C]	-767.1	-709.1	-713.1	-698.8	-725.6	-726.9	-742.3	-721.1	-734.9	-722.8
21	describe and explain various types of light emissions (e.g., chemiluminescence, bioluminescence, incandescence, fluorescence, phosphorescence, triboluminescence; from an electric discharge or lightemitting diode [LED])	-679.2	-679.8	-679.6	-726.9	-679	-736	-679.5	-702.6	-670.1	-680
22	identify and label the visible and invisible regions of the electromagnetic spectrum	-800	-800	-800	-800	-800	-800	-800	-800	-800	-800
23	describe, on the basis of observation, the characteristics and positions of images formed by plane and curved mirrors (e.g., location, orientation, size, type) , with the aid of ray diagrams and algebraic equations, where appropriate	-693.4	-692.6	-687.8	-743	-705.1	-719.7	-750.5	-691.4	-723.4	-647.5
24	explain the conditions required for partial reflection/refraction and for total internal reflection in lenses, and describe the reflection/refraction using labelled ray diagrams	-740.9	-689.8	-654.3	-654.5	-617.9	-741.8	-676.6	-646.5	-682.6	-621.6
25	describe the characteristics and positions of images formed by converging lenses (e.g., orientation, size, type), with the aid of ray diagrams	-651.3	-625.5	-595	-621.7	-652	-705.2	-708.5	-625.4	-643.7	-565.7
26	identify ways in which the properties of mirrors and lenses (both converging and diverging) determine their use in optical instruments (e.g.,cameras, telescopes, binoculars, microscopes)	-713.5	-756.5	-623.9	-617	-671.7	-693.2	-715.6	-703.2	-675.9	-702.2
27	identify the factors, in qualitative and quantitative terms, that affect the refraction of light as it passes from one medium to another	-720.3	-671.2	-738.6	-720.3	-754.2	-753.8	-711.7	-695.2	-753.4	-701.3
28	describe properties of light, and use them to explain naturally occurring optical phenomena (e.g.,apparent depth, shimmering, a mirage, a rainbow)	-762.5	-762.1	-761.1	-730.8	-792.5	-760.6	-787.6	-762.2	-764.3	-765.1

Figure 2 – b, Curriculum Coverage of Grade 10

7. Conclusion

In this work, we explored the use of machine learning techniques, in particular, probabilistic topic models, in assisting education practitioners to analyze online discussion data. Our methodology is to decompose a large corpus of textual materials collected from online learning platforms into distinct and semantically meaningful clusters (i.e., topics). Representing documents according to their topic relevance can greatly facilitate the query, organization and comparison of a large corpus. More importantly, the recovered topics can be used by practitioners to map the students' learning performance to the instructor's learning objectives, via temporal, interpretative and comparative analyses.

Acknowledgement. Siwei Lyu is supported by the National Science Foundation under Grant Nos. IIS-0953373 and CCF-1319800. Jianwei Zhang is supported by the National Science Foundation through Grant No. 1122573.

References

[1] 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.

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

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

[4] Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2004). Integrating topics and syntax. In *Advances in neural information processing systems* (pp. 537-544).

[5] Hofmann, T. (1999, August). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 50-57). ACM.

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

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

[8] Ming, N., & Ming, V. (2012). Predicting student outcomes from unstructured data. In *UMAP Workshops*.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.*

[14] 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.

[15] 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.

[16] 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.

[17] 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*.