Automatic extraction of interpretable topics from online discourse

Yonghe Zhang, Beijing Normal University, China, yonghe617@gmail.com
Nancy Law, The University of Hong Kong, Pokfulam, Hong Kong, nlaw@hku.hk
Yanyan Li, Beijing Normal University, China, liyy1114@gmail.com
Ronghuai Huang[*], Beijing Normal University, China, huangrh@bnu.edu.cn

Abstract: Teachers adopting CSCL often face the challenge of handling massive textual information, and finding it difficult to have a clear grasp of the topics being addressed in the discourse. Topic modeling, an emerging field in machine learning, has the potential to solve this problem by automatically extracting from text collections formal representations of latent topics. However, the interpretation of latent topics is still a challenge, which hinders the use of this state-of-the-art technology from wider use in CSCL contexts. In a recent paper, we put forward a novel topic discovery method, the fLDA model, based on Minsky's Frame theory. This method has the advantage of providing outputs that are potentially more easily interpretable for generating the topic of each thematic cluster. In this paper, we show how fLDA can be used in extracting and visualizing the topics of asynchronous online discourse from four classrooms.

Introduction

Online discussion forums provide students with the opportunity to explore different problems and topics through discussion in an unconstrained manner, allowing more student-centered interactions to take place. It also provides a record of the explorations that teachers can use to gain an understanding of the students' concerns and on that basis to make facilitation decisions. However, making sense of the massive amounts of text in the posted messages is a daunting challenge for teachers adopting CSCL in their teaching repertoire. How can a teacher readily find out the key topics of discussion among hundreds of posted messages? There is a need for semantic tools that can identify key topics in online discussions to support pedagogical decision-making.

Summarizing massive free style text has been a long-standing issue in computer science, esp. in machine learning. One of the most relevant research areas to tackle this problem is *text clustering* (also referred to as *document clustering*). Text clustering methods can be used to automatically group documents into a list of clusters. Each cluster is regarded as a collection of documents on a latent topic. By interpreting latent topics and examining their distribution, users can gain an overview of the content focus of a collection of texts.

Recent research in text clustering shows that the interpretability of latent topics is still a challenge (Blei, in press). Earlier text clustering methods based on the Vector Space Model (Salton, Wong, and Yang, 1975) using vector similarity measures (e.g. cosine similarity) do not provide further semantic clues for the interpretation of topics. Latent Semantic Analysis (Deerwester, Dumais, Landauer, Furnas, Harshman, 1990) uses a set of latent variables to represent topics, but still fails to provide intuitive interpretation for each topic. Topic models such as LDA (Blei, Ng and Jordan, 2003) use a set of weighted words to represent each topic. These have been used successfully to identify research topics among tens of thousands of scientific articles (Griffiths and Steyvers, 2004). Other work related to the interpretability of topic models include coherence measures of topics manually (Chang, Boyd-Graber, Gerris, Wang, and Blei, 2009), automatically (Newman, Han-Lau, Grieser and Baldwin, 2010; Musat, Velcin, Trausan-Matu, and Rizoiu, 2011), interpretability improvement by semi-supervised method (Zheng, 2008) and connecting topics, e.g. hLDA (Blei, Griffiths, Jordan and Tenenbaum, 2003), PAM(Wei and Andrew, 2006), hPAM (David,Wei and Andrew, 2007).

This paper introduces our recent work on a novel topic discovery method, the fLDA model (Zhang, Li and Huang, under review), and presents our preliminary exploration on using this to identify and interpret the online discourse data collected from four classrooms that used Knowledge Forum® to support students' learning.

Topic Models and the Interpretability Challenge

Probabilistic topic models are a suite of algorithms whose aim is to discover the hidden thematic structure in large archives of documents (Blei, in press). The most famous topic model is Latent Dirichlet Allocation (LDA) (Blei et al, 2003). The intuition behind LDA is that documents exhibit multiple topics. A *topic* is formally defined as a distribution over a fixed vocabulary. For example, a document on the topic of *education* would have a high probability of containing words about pedagogy and learning while one on the topic of *computer*

[*] Corresponding author is Ronghuai Huang, huangrh@bnu.edu.cn.

This paper is supported by the Chinese National Natural Science Fund (61075048) and the Information Technology Strategic Research Theme of the University of Hong Kong.

science would likely contains words about hardware and software. Technically, LDA assumes that topics are specified before any document is generated. Each document in a collection is regarded as generated in a two-stage process (document generative process):

1. Randomly choose a distribution over topics.
2. For each word in the document
   (a) Randomly choose a topic from the distribution over topics in step #1.
   (b) Randomly choose a word from the corresponding distribution over the vocabulary.

Based on this assumption, the goal of topic modeling is to automatically discover the topics from a collection of documents. The documents themselves are treated as observed data, while the topic structure—the topics, per-document topic distribution, and the per-document per-word topic assignment—is the *hidden structure*. The central computational problem for topic modeling is to use the observed documents to infer the hidden topic structure, and hence is a "reverse generative process.

Topic modeling algorithms generally fall into to two categories—sampling-based algorithms (Steyvers and Griffiths, 2006) and variational algorithms (Wainwright and Jordan, 2008). For a good discussion of the merits and drawbacks of both, see Asuncion, Welling, Smyth and Teh (2009). Topic modeling is an emerging field in machine learning. Among the many challenges in this field, the most basic one may be the evaluation and model checking problem—how to evaluate a topic model by the meaningfulness of the topics identified (Blei, in press). This problem has led to many efforts on evaluation and improvement of topic interpretability.

Current topic interpretability research focuses on evaluation of topic interpretability and interpretability improvement. To evaluate topic interpretability, popular strategies include the use of coherence measure of topics using a qualitative approach (Chang et al, 2009) and a lexicon-based automatic approach (Newman et al, 2010; Musat et al, 2011). To improve topic interpretability, many efforts have focused on discovering topic hierarchy. Ning (2008) incorporates human intervention in the topic modeling process, such as shaping topics with human knowledge. Current topic models use a set of weighted words to represent a topic.

The fLDA Model

Our work focuses on improving topic interpretability by representing each topic in a more human-readable form. To improve user interpretability of topics, we need to understand how humans achieve understanding of situations represented by words. There are many theories in discourse psychology that try to explain text comprehension, e.g. Frame Theory (Minsky, 1975), Script Theory (Schank, 1975), Story Grammar (Rumelhart, 1975) and propositional representation of discourse (Kintsch and van Dijk, 1978). Among these theories, Minsky's Frame Theory can be more easily implemented in artificial intelligence applications in terms of its data structure.

According to Minsky's (1975) Frame Theory (or Schema Theory), when one encounters a new situation or makes a substantial change in one's view of the situation, one selects from memory a structure, called a frame. A frame is a data-structure for representing a stereotypical situation, like being in some ordinary living spaces or performing certain activities. From a computational perspective, a frame is a set of slot-value pairs. Slots are stable for a frame while the value for each slot is adaptively assigned to represent each specific situation. For example, a story frame may have slots named time, place, thing and event. So a theft story may be: At 8:00 pm inside a jewelry store, a diamond watch was stolen when the shop-keeper was distracted by a mob shouting outside. Then a frame for this story can be written as (time="8:00 pm", place="a jewelry store", things="a diamond watch", shop-keeper, a mob shouting, event="distracted, stolen"). The level of detail in value assignment of frame slots is assumed to reflect the extent of one's understanding of the situation. For example, event="the shop-keeper was **distracted** by a mob, a diamond watch was **stolen**" is more understandable than event="distracted, stolen". The Frame Theory has inspired much productive research in cognition (Solso, MacLin and MacLin, 2004). It is apparent that the externalization of topics in the form of frames helps one to understand the topics. Hence in this study we explore the use of a frame-based method to extract topics from document collections in order to find out whether the output is more readily interpretable as meaningful topics by human readers. In the remainder of this section, we will briefly describe the frame-based topic discovery model we developed based on the LDA model, named fLDA (Zhang et al, under review).

Definitions of key terms in the fLDA model

The following are the definitions of some basic terms used in the fLDA model.

Word—A *word w* is the basic unit in text, and is defined as a string in its original form.

Document—A *document d* is a sequence of words extracted from the text. It can be denoted as d = {w_1,w_2, . . . , w_n}. The *i-th* element of *d* is referred to as the *i-th word token* or *i-th token,* which is conceptually different from the term *word*.

Corpus—A *corpus c* is a collection of documents.

Term frequency—The term frequency of a word *w* in corpus *c* is the number of occurrences of w in all documents contained in *c*.

Co-occurrence frequency—The co-occurrence frequency of word w_1 and w_2 in corpus c is the number of occurrences of w_1 in any document containing w_2 in *c*.

Topic frame—A *topic frame f* is a quadruple of semantic slots: focus, features, events and related things. A ***focus*** is a single word representing an entity, and can be regarded as a ***central concept*** of a topic. A ***feature*** is a weighted word representing an entity property. Each frame contains a set of features. An ***event*** is a weighted word representing some action. A ***related thing*** is a weighted word representing entities. Each of the features, events and related things is weighted by its co-occurrence frequency with respect to the focus of the frame, for frames having non-zero weight. It should be noted that ontologically the focus words are also things. Features, events and related things can be regarded as ***foil concepts*** or ***foils*** to the central concept. A "foil" is used in text comprehension studies to refer to a concept associated with a central concept. In this study, a frame is make up of a focus and its foils, and a foil of a focus can be a feature, an event or a thing related to the focus.

Topic—A *topic T* of corpus *c* is a set of topic frames and their weights pair, and these topic frames reflect the main content of the topic-related documents. It can be denoted as $T = ((f_1,pf_1), (f_2,pf_2), \ldots , (f_t ,pf_t))$, where pf_i is equal to the term frequency of the focus of f_i.

Input and output of fLDA topic modeling algorithm

The topic modeling approach of fLDA is to input a corpus with POS (Part-Of-Speech) tags for each word, and output a set of topics defined by the previous sub-section. POS-tagging tools are available online in several languages, e.g. the Stanford POS tag tool for English and the ICTCLAS tool for Chinese.

In the fLDA model, each word has a semantic class (one of “thing”, “feature”, “event” or “other”) and a topic. fLDA assumes a similar corpus generative process as LDA. Based on this assumption, we have developed a Gibbs sampling-based algorithm to assign a semantic class and a topic for each word. In Zhang, Li and Huang (2011), we use a pseudo dataset generated using the corpus generative process to test the recall rate and performance of this algorithm. Intuitively, every 10~30 new documents stay on one topic, and each word token in a document has 50% probability to be on the document topic and 50% to be of other topics randomly. Due to the lack of a fully trusted quality indicator for the topic-modeling algorithm, we present a visualization of the topic discovery process. This visualization is generated by the algorithm in JPG image format. Due to space limitations, only a part of the visualization is shown in figure 1. The numbers at the top of the images indicate the iteration number and each image represents the matrix for the document-topics. Each row of an image represents a document, and each column a topic. The darkness of a pixel indicates the topic word frequency of a document on a topic. Tiny horizontal lines are used to indicate the boundary of every 10 documents. At iteration zero, the color of blocks in the image is uniformly distributed, as it is assumed that every document has similar word frequency on topics. As the number of iteration increases, four vertical line segments become more and more clear. Each of these segments is made up by several adjoining blocks in the same column. This indicates that the corresponding documents have high word frequency in the corresponding topic of the relevant blocks. The documents with darkened blocks in the same column relates to the same topic. The image of iteration 400 shows a clear topic distribution, and this distribution is consistent with the parameters predefined in the data generation rule.

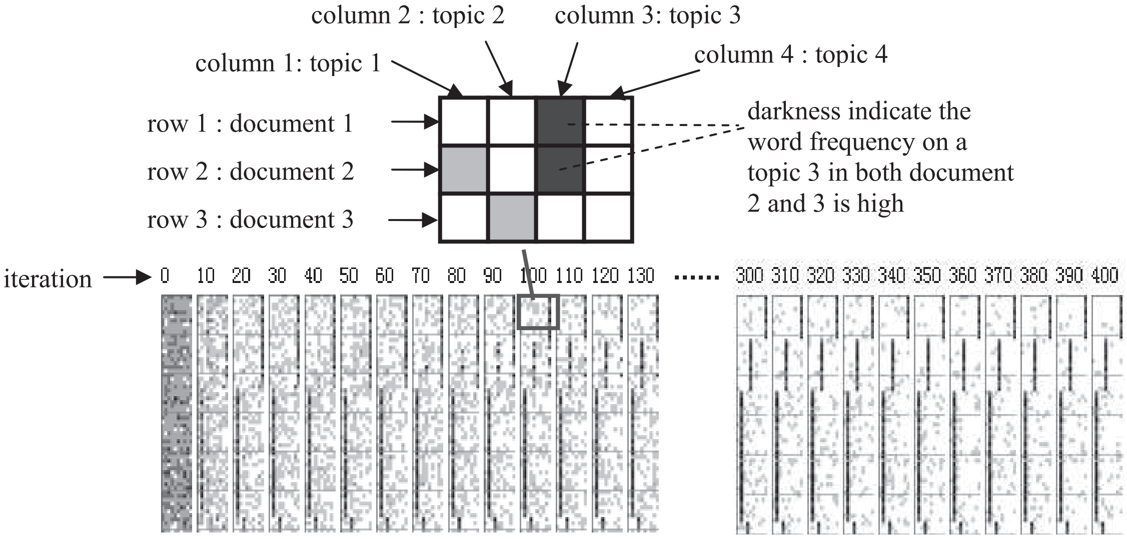

Figure 1. Visualization of the topic discovery process by the fLDA algorithm on a pseudo dataset.

We chose one of the four Knowledge Forum® discourse datasets used in this study to test the quality of tagging on the semantic classes, which is the basis of frame slots using this algorithm (the 6B dataset referred to in the next section), and used Cohen's Kappa (Jean, 1996) as the cross-validation indicator. Two graduate students achieved kappa=0.73 for their tagging, while the algorithm-human kappa=0.69. Although this is far from perfect, the result can be regarded as an acceptable one.

Interpretability of fLDA Generated Topic Outputs Using Authentic CSCL Data

To explore whether fLDA can help one to interpret the topic foci of online discourse in CSCL settings, we analyzed the discourse corpuses from four primary school Chinese Language classrooms participating in the Knowledge Building Teacher Network led by HKU-CITE (http://kbtn.cite.hku.hk). Some basic information about these corpuses is shown in Table 1. A *note* is a message posted by one or more students. The four classrooms belong to two schools: 6B, 6C belong to one school and 5A, 5B belong to another. The classrooms from the same school shared the same discussion topic based on the set school curriculum. The theme "Hong Kong (HK) Kids" was centered around negative media reporting about Hong Kong children's lack of general life skills and inability to handle adversity. "Ice duck" is the title of a passage in the Chinese Language textbook, which is a story about the pursuit of one's ideal in life.

Table 1: Basic information of the CSCL discourse corpuses selected for this study.

Class	Number of notes generated	Discussion duration (weeks)	Theme of discussion
6B	274	10	HK kids
6C	76	8	HK kids
5A	167	18	Ice duck
5B	199	15	Ice duck

Topic extraction

In order to be able to compare the topic frequencies in the discussions generated from different classrooms sharing the same theme, we combined the corpuses of 6B and 6C in one analysis, and the corpuses from 5A and 5B in another. fLDA is an exploratory method and the user needs to pre-set the number of topics to be extracted. In our preliminary exploration, analyses using pre-set topic numbers from 4 to 10 were conducted. We find that outputs from too few or too many pre-set topics are more difficult to make meaningful interpretations. In this study, the topic number used is 5. In all runs of the algorithm, the number of iterations was set to 1000.

In this paper, we focus our reporting on the 6B_6C corpus. The theme of the discussion is "Hong Kong Kids". The output generated by the fLDA algorithm is reprocessed to show the top 3 weighted frames of each topic and the three documents that are most representative for the topic. Table 2 presents the high co-occurrence words for the top three frames generated by fLDA for each of the five topics. The original text of the students' discourse was in Chinese. The English translation of the words are provided in brackets.

The meaning of each topic is not immediately evident based only on the frame words in Table 2. The following steps are taken to formulate an understanding of the topics:

1. For each focus word in a topic, identify the co-occurring words in the other frames and rank order them in decreasing frequency.
2. For the high co-occurrence pairs, locate the text segments containing these pairs to understand the relationships expressed in the documents on these pairs.
3. Synthesize the meanings of the co-occurring words in the top three frames of each topic to generate the meaning of each of the five identified topics.

To illustrate how we can follow the above steps to arrive at the meaning of a topic, we use the first topic in Table 2 as an example.

Table 2: High co-occurrence content words in frames of Topic 1 from a five-topic analysis of the 6B_6C corpus (English translation of words and frequency of occurrence in brackets).

Topic	Focus (things)	Feature	Event	Related things
1	制度 (system, 22)	难(difficult, 2)	学(learn,4), 教育(education,4) 出现(occur,3), 减少(reduce, 3)	教育 (education, 13), 香港(Hong Kong, 8),精英制度(elitism, 7)
	香港 (HK, 19)	穷(poverty, 2)	学(learn, 4), 出现(occur,3), 蔓延(spread,3) 照顾(take care, 2)	制度(system, 8), 精英制度 (elitism, 5) 教育(education, 4), 国家(country, 3)
	教育 (education, 12)	难(difficult, 2)	学(learn, 4), 出现(occur, 3), 入学(admission, 2)	制度(system, 13), 精英(elite, 6), 香港(Hong Kong, 4), 父母(parent, 3), 政府(government, 3)
	佣人 (maid, 13)		遇到(encounter ,4), 帮助 (help ,4), 培训(train ,3), 听 (listen,3),独立(independent,2)	培训佣人(training of maid, 4), 儿童(child, 2), 政府(government, 2), 方法(method, 2)

First, as we can see in Table 2, the foci of topic 1 are “system”, “Hong Kong” and “education”. According to our contextual knowledge and the high co-occurrence of these three words in the topic frames, we predict this topic is about the relation between Hong Kong’s education system and the cause of the “HK Kids” phenomenon (which refers to the children’s lack of general life skills and inability to handle adversity).

In the second step, we find that among the word pairs containing the foci words, the words with the highest co-occurrence frequency are “elitism” and “elite”. We locate the text segments containing these pairs as bellow. The English translation of some of these segments are listed below:

- Prof. Chen point out that Hong Kong’s education system lead to elitism.
- Since the whole society is focused on fostering elites, parents in our education system wants to send their kids to prestigious schools. Otherwise, it will be difficult for these kids to survive in an elitist society.
- If the Hong Kong education system does not encourage elitism, will the “Hong Kong Kids” phenomenon gradually disappear?
- “Hong Kong Kids” may be caused by the elitism oriented education system in Hong Kong.

It is clear from these text segments that there is a strong semantic overlap across the 4 text segments and it is not controversial to identify that topic 1 is about the Hong Kong elitist education system being the root cause of the “HK Kids” phenomenon. Table 3 presents the interpretations we arrived at on the five topics generated by fLDA. The topic words found in the frames are highlighted in colors according to their semantic classes (things in blue, features in green, events in red).

Table 3: Interpretation of the topics identified from the_6B_6C discourse corpus.

Topic	Interpretation of topic
1	诱发港孩出现的原因是因为整个社会过分重视培育精英，于是教育制度令父母催谷子女入学名校，家长只是注重成绩而忽略培育个人品格和独立能力 。(The HK kids syndrome occur because of the elitist education system. Parents drill children for admission to famous schools. Parents pay too much attention to children’s school performance, ignoring the development of children’s integrity and independence.)
2	「三低」特征包括自理能力低、情绪智商低、抗逆力低。生活小节均由他人代劳，不懂得照顾自己，待人接物能力差。(“3 low” means low self-care capacity, low EQ and low resilience to adversity. HK kids are helped by others in their daily life, not able to take care of themselves, with poor socialization capacity.)
3	有些香港孩子有情绪病，是父母过度保护孩子引致。他们介入孩子之间的争吵、深怕子女吃亏，经常代子女出头, 令孩子变得脆弱不堪。(Some HK children have mood disorders caused by overly protective parents who interfere with their children’s disputes with others, afraid of their kids being disadvantaged and always fighting on their kid’s behalf, causing their children to be unable to handle any adversity.)
4	父母应以身作则，对自己的行为和观念反思，尽早放手让孩子学习, 放心让孩子撞板，令孩子多嚐试失败的滋味，学习解决问题的能力。(Parents should lead by example, and reflect on their own behavior and ideas, release their control as early as possible to let the children learn and experience failure so that they can build the capacity to solve problems.)
5	家长聘请佣人照顾子女，使他们变得事事依赖，应该教导家长去培训子女的独立能力。(Parents hire maids to look after their children, making them rely on others for everything. Parents should be taught to train their children’s capacity to be independent.)

Towards the end of this unit, the teacher asked the students to summarize their discussions. The content of the students’ summaries can be used as a reference for validation of our interpretation of the topics. The following is a translation of the students’ discussion summaries:

- First, we think the HK Kids’ parents would not try to remedy their ways of disciplining their children because they think sending their children to prestigious schools and making them study all the time is correct, and do not pay attention to the real needs of their children.
- Second, we do not think that “HK Kids” are caused by the government enforcing the education system. Maybe parents are too busy with their work, and want to compensate for not giving enough love to their children by giving them more material comforts, and hiring maids to look after them. However, in this way, children cannot learn to become independent and they are used to rely on others to solve their problems. Some students consider the HK education system as fostering elitism which cause HK Kids
- Third, we think that in order to solve HK Kids problem, parents should pay attention to their children's habits and train them to become independent starting from childhood. The government should offer courses for children to improve their self-understanding, and to train them to manage their own emotions.

- Fourth, the emergence of "Hong Kong Kids" is not a regional problem. Because of fast-paced lifestyle in Hong Kong, many parents do not have time to take care of their children. Another reason may be related to economic conditions. HK parents can employ maids to take care of their children leading to children's weak problem-solve skill. But in mainland China, children in poor families have to take care of themselves from childhood.)

The first part of the above students' summaries is similar to our interpretation of topic 1: parents over-emphasizes children's school education performance too much, and overlook the training of their childred's independence. However, in this part, students do not mention elitism.

The second part of the summary is similar to our interpretation of topic 5: parents employ maids to take care of their children, resulting in their weak ability to be independent. Here, elitism is mentioned as a factor associated with the HK Kid syndrome, though not as central as picked up by the fLDA algorithm.

The third part is similar to topic 4: parents should foster children's independence. The children's summary is written in more concrete terms than the interpretation presented in Table 3.

The final part of the students' summary is similar to topic 5: Due to the better economic conditions in HK, parents are able to employ maids to take care of their children. The content of this summary is also related to topic 1: in the frames of topic 1, the words "poverty" and "countries" turn out to be mentioned when comparing HK with other regions.

Through this comparison, we find that our interpretation can cover most of the ideas students expressed in their own summary of discussions. On the other hand, we cannot find summaries with content similar to topics 2 and 3. So our approach may have potential in helping readers identify and interpret topics.

Topic trend analysis

One use of topic analyses is to compare the discussion topics across corpuses and over time. To do so, we first tag each note with relevant topics. A note is tagged with a specific topic if it contains at least 2 topic words of the topic. It is possible for a note to have all five or none of the identified topics tagged. Then we create for each corpus a visualization, called per-week topic trend graph, to show the number of notes containing each of the topics for each week. As shown in figure 2, this visualization of the 6B corpus reveals that the number of notes change significantly over time. In week 2, there are peaks for all topics. Afterwards, all topics have a low occurrence except for the "none-of-the-topics" category, which had another peak in week 6.

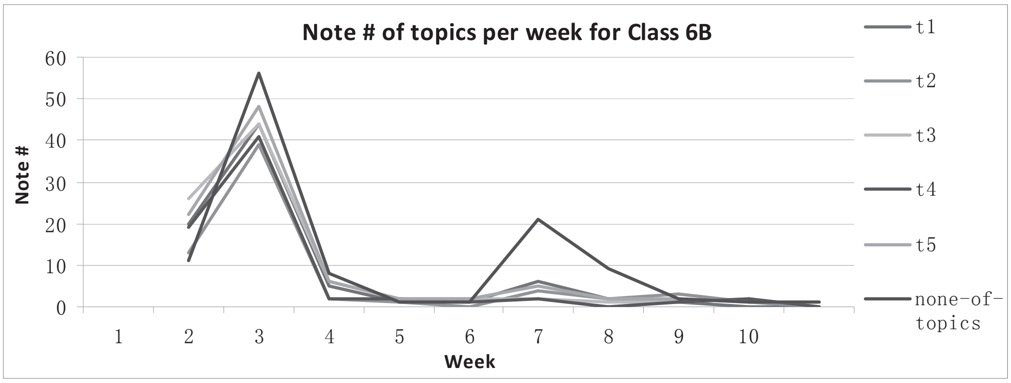

Figure 2. Per-week topic trend graph of the 6B corpus.

However, per-week topic trend graph of class 6B cannot reveal clear changes in topical focus when the level of discourse activity changes significantly over time. To provide a better representation of the change in topic focus in terms of the concerns expressed in notes, we make another visualization, per-N-notes topic trend graph, based on grouping every N=20 notes in order of created time. Per 20-notes topic trend graphs for all datasets are shown from figure 3 to 5.

As shown in figure 3, the wave shape curves in the visualization reveal note numbers of topics have declined and increased for more than two times. By looking at the curves within first 60 notes, we find that topic 3 is more stable than other topics. This may indicate that the 6B students focus more on family problems than other causes of Hong Kong kids. Another interesting finding can be recognized at the point of 121-140 notes where the note number of none-of-the-topics category suddenly increases while other categories reduce a lot, indicating possibly some new topics may have emerged.

Figure 4 shows the topic trend graph for the 6C corpus, which is quite different from that for the 6B corpus. From the point of 20-40 notes to the one of 40-60 notes, the note number of the "none" category goes down while the numbers of topical notes increase. This indicates different topical foci for the classes 6B and 6C in their discussions on the same theme.

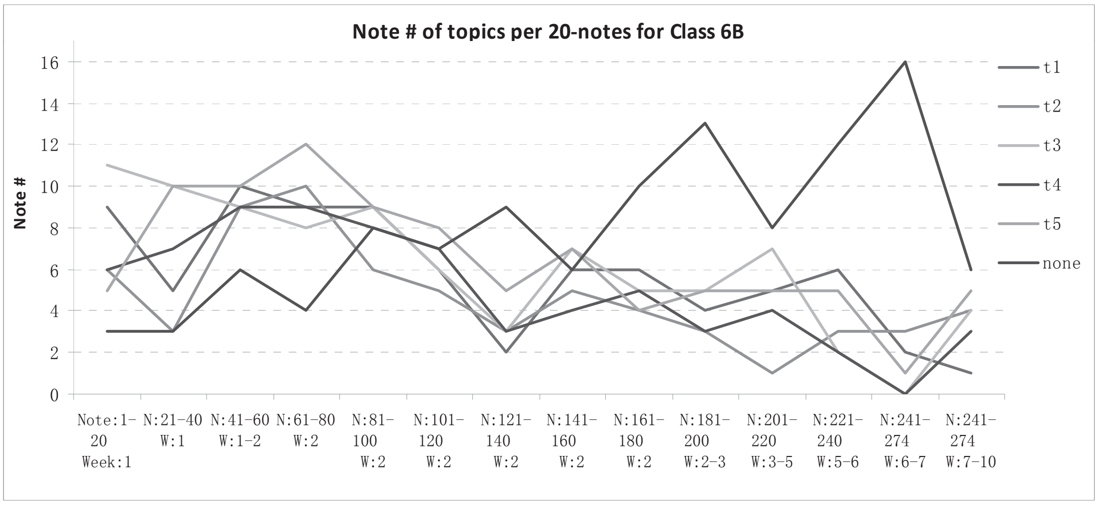

Figure 3. Per-week topic trend graph of 6B data set.

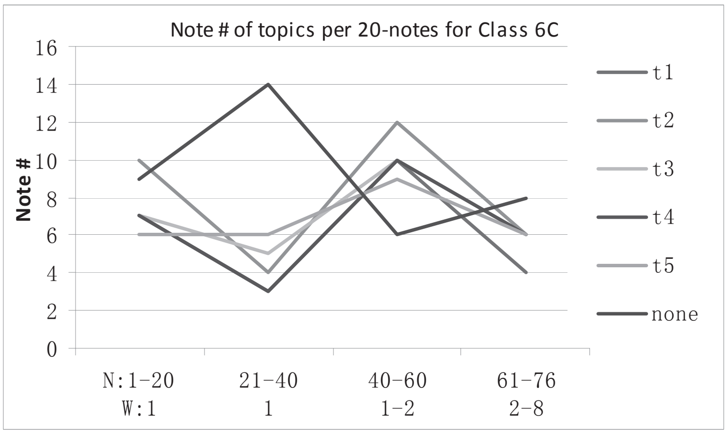

Figure 4. Per-week topic trend graph of the 6C corpus.

We analyzed another set of corpus from a pair of classrooms from another school, 5A and 5B, and compare the topic trends between them based on their per-20-notes topic trend graphs, as shown in figure 5. It seems 5A had very few notes related topic 4 and 5. In contrast, 6B had very few notes on topics 1, 2 and 3. The ways the discourse developed in terms of the major topics are different as well. For 5A, topic 1 became a hot topic in the stage of 1-80 notes, then "cooled down" slowly afterwards, and finally disappeared. For 5B, topic 4 become a hot topic in the stage of 1-40 notes, became a rare one in the 41-80 stage, and then become a normal topic found in 25% of the notes in the later stage.

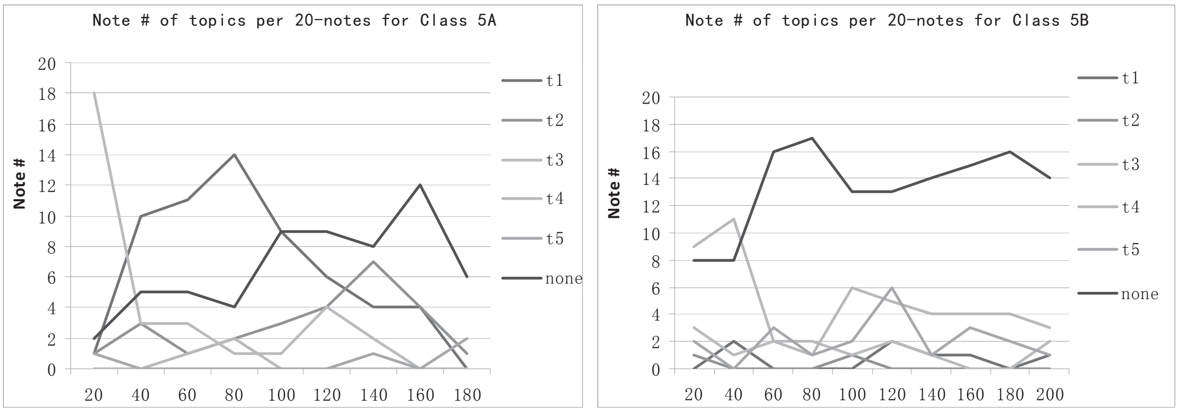

Figure 5. Per-week topic trend graph of the 5A (left) and 5B(right) corpuses.

Conclusion

We have demonstrated two promising applications of fLDA in analyzing CSCL discourse corpuses: topic extraction and topic trend analysis. Topic extraction starts with performing the fLDA algorithm to generate frames for topics from a collection of notes. Notes can be tagged with different topics according to topic word frequency. Based on these frames, we can make meaningful interpretations for each topic through a few steps. Topic trend analysis can be done with per-N-notes topic trend visualization. Interpretation of this kind of visualization can potentially be used to help teachers to identify the focal concerns in students' online discussions. More work involving the participation of teachers in evaluating the validity and usefulness of such analyses is necessary to determine the value of such a topic discovery and visualization tool.

References

Asuncion, A., Welling,M., Smyth, P. & Teh, Y. (2009). On smoothing and inference for topic models. *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, p.27-34, June 18-21, 2009, Montreal, Quebec, Canada.

Blei, D. (in press). Introduction to probabilistic topic models. *Communications of the ACM*, Retrieved March 17, 2012, from http://www.cs.princeton.edu/~blei/papers/Blei2011.pdf.

Blei, D., Ng, A., & Jordan,M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993-1022.

Blei, D., Griffiths, T., Jordan, M. & Tenenbaum, J. (2003). Hierarchical topic models and the nested Chinese restaurant process. *Advances in Neural Information Processing Systems 16*, Cambridge, Retrieved March 17, 2012, from http://books.nips.cc/papers/files/nips16/NIPS2003_AA03.pdf.

Chang, J., Boyd-Graber,J., Gerris,S., Wang,C. & Blei, D. (2009). Reading tea leaves: How humans interpret topic models . *Advances in Neural Information Processing Systems 22*, Retrieved March 17, 2012, from http://books.nips.cc/papers/files/nips22/NIPS2009_0125.pdf .

David, M., Wei, L. & Andrew M. (2007). Mixtures of hierarchical topics with Pachinko allocation, *Proceedings of the 24th International Conference on Machine Learning*, p.633-640, June 20-24, Corvalis, Oregon.

Deerwester, S., Dumais, S., Landauer, T., Furnas, G. & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6): 391-407.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Science*, 101, 5228-5235.

Jean, C. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2), 249–254.

Kintsch, W. & van Dijk, T. A. (1978). Toward a model of text comprehension and production, *Psychological Review*, 85, 363-394.

Minsky, M. (1975). A framework for representing knowledge. In P.Winston (Ed.), *The Psychology of Computer Vision*. New York: McGraw-Hill, pp. 211-277.

Musat, C. C., Velcin, J., Trausan-Matu, S. & Rizoiu, M. A. (2011). Improving topic evaluation using conceptual knowledge. *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, Retrieved March 17, 2012, http://www.aaai.org/ocs/index.php/IJCAI/IJCAI11/paper/view/3010/3754.

Newman,D., Han-Lau,J., Grieser, K. & Baldwin, T. (2010). Automatic evaluation of topic coherence. *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 100-108. June 2-4, 2010, Los Angeles, California, USA..

Rumelhart, D.E. (1975). Notes on a schema for stories. In: Bobrow D.G. & Collins, A.M. (Eds.) *Representation and understanding: Studies in cognitive science*, pp. 211-236. New York: Academic Press.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18 (11), 613-620.

Schank, R.C. (1975). *Conceptual information processing*. Amsterdam: North-Holland.

Solso,R. L., MacLin, M. K., MacLin, O. H. (2004). *Cognitive Psychology* (7th Edition) . Allyn & Bacon.

Steyvers, M. & Griffiths, T.(2006). Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis & W. Kintsch, (Eds.), *Latent Semantic Analysis: A Road to Meaning*. Lawrence Erlbaum.

Wainwright, M. & Jordan, M. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2), 1-305, December 2008.

Wei, L. & Andrew, M. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. *Proceedings of the 23rd International Conference on Machine Learning*, pp.577-584, June 25-29, Pittsburgh, Penn.

Zhang, Y., Li, Y. & Huang, R. (under review). fLDA: A Topic Model based on Frame Representation.

Zheng, N. (2008). *Discovering interpretable topics in free-style text: diagnostics, rare topics, and topic supervision*. Doctoral dissertation, The Ohio State University, USA.

Acknowledgments

This study is supported by funding from the Information Technology Strategic Research Theme of the University of Hong Kong and the Chinese National Natural Science Fund.