Topic Facet Modeling: Semantic Visual Analytics for Online Discussion Forums

I-Han Hsiao
School of Computing, Informatics & Decision Systems Engineering,
Arizona State University, Tempe, AZ, USA
+1 480 727 5466
Sharon.Hsiao@asu.edu

Piyush Awasthi
School of Computing, Informatics & Decision Systems Engineering,
Arizona State University, Tempe, AZ, USA
pawasth1@asu.edu

ABSTRACT
In this paper, we propose a novel Topic Facet Model (TFM), a probabilistic topic model that assumes all words in single sentence are generated from one topic facet. The model is applied to automatically extract forum posts semantics for uncovering the content latent structures. We further prototype a visual analytics interface to present online discussion forum semantics. We hypothesize that the semantic modeling through analytics on open online discussion forums can help users examine the post content by viewing the summarized topic facets. Our preliminary results demonstrated that TFM can be a promising method to extract topic specificity from conversational and relatively short texts in online programming discussion forums.

Categories and Subject Descriptors
K.3.2 [Computer and Information Science Education]: Computer Science Education.

General Terms

Keywords
Learning Analytics, Automated Assessment, SLDA, LDA, TFM, Topic Modeling, Programming, Discourse Analytics, Discussion Forums.

1. INTRODUCTION
Forums or discussion boards are popular trouble-shooting/problem-solving technologies for online courses. In current programming language learning, students are exposed to a lot of help in preparing code (readings, intelligent tutorials, worked examples etc.), as well as in-situ support during coding exercises (e.g., debuggers, unit tests, dynamic code examples, dialogue tutor etc.) (Barnes & Stamper, 2008; Boyer et al., 2011; Brandt, Dontcheva, Weskamp, & Klemmer, 2010; Brusilovsky, 2001; Pirolli & Anderson, 1985) Free online discussion sites (i.e. stackoverflow (http://stackoverflow.com), Dream.In.Code (http://www.dreamincode.net), etc.) allow programmers and learners to reach out for help so that they can freely discuss programming problems, ranging from general to specific and simple to complex topics. These sites therefore not only throw open unbounded topics in the form of questions and answers, but are especially attractive for open-ended problem discussions. In online open corpus discussion forums, the scale and types of posts are often very diverse in terms of user background, coverage of topics, post volumes, post-response turnaround rates, etc. In addition, these platforms are usually not moderated or guided by teachers or teaching assistants, but are essentially governed by the community. There has been considerable research on strategies to filter the quality of content and encourage participation of online communities via crowdsourcing, rating, tagging, badges, etc. (Hsiao & Brusilovsky, 2011; Jeon, Croft, & Lee, 2005; Kittur, Chi, & Suh, 2008; Snow, O’Connor, Jurafsky, & Ng, 2008). Such social mechanisms tend to filter and point out the most possible correct solutions. However, in the context of online learning, the correct solutions may not necessarily be the best next steps for all learners (Graesser, VanLehn, Rose, Jordan, & Harter, 2001; van de Sande & Leinhard, 2007). Without personalized learning support, the voted best solutions can be easily too sophisticated or overly generic. With rapid growth in content and communities, therefore, comes the need for more personalized and intelligent support.

Currently, most forums or discussion boards lack dynamic and extensive content analysis, mainly due to the discussion content is often open-ended and filled with nebulous semantic structures. It is so-called ill-defined problems, which have usually been within the reach of qualitative human-coded methods, and are typically difficult to scale. Moreover, these methods are challenging to keep persistent traces for estimating students’ current knowledge (Blikstein, 2011). Due to the increasing volume and run-time calculation complexities (i.e. exhaustive linguistic features in natural languages), most of the discourse-centric in-depth analyses were performed offline and the lesson learned could only be applied in the next iteration of system development or suffering from shallow analytics (i.e. participation & contribution). Recently, however, we have begun to see some studies that focus on dynamic support for users, such as Cohere – structured online discourse and summary (Shum, 2008); Social learning analytics (Shum & Ferguson, 2012); Asynchronous conversation analytics (Hoque, Caremini, & Joty, 2014). Yet, there has been no conclusive or comprehensive technological support as well as systematic studies to date on large-scale discussion forums. In addressing the research issues discussed above, this paper focuses on studying viable algorithmic technologies that support large-scale discussion forums dynamically, specifically targeted at programming discussion forums. We propose a novel Topic Facet Model (TFM) to extract forum posts semantics for uncovering the latent structural topics. We further prototype a visual analytics interface to present online discussion forum semantics. We
assume users arrive a programming discussions site to seek for help or to offer help, but not all of them are domain experts. Therefore, we hypothesize that an intelligent support through analytics on open online discussion forums can help users examine the post content by viewing the summarized topic facets.

The rest of the paper is structured as following: in section 2, we review a series of related work regarding to discourse-centric learning analytics and topic modeling. In section 3 & 4, we describe the novel generative Topic Facet Model and illustrate a learning analytic prototype. Lastly, we discuss data collection and evaluation results.

2. LITERATURE REVIEW

In order to address learner-centered technology for large-scale online learning, Learning Analytics (Siemens & Baker, 2012) has demonstrated promising results, especially in interdisciplinary convergence. For example, the Signals project at Purdue University is one of the pioneering examples of the successful application of academic analytics that integrate predictive modeling and report significantly higher grades and retention rates than were observed in control groups (Arnold, 2010). Nonetheless, the majority of learning analytics discuss visual representations or the system’s usefulness while the core should be focused on real impact to improve learning or teaching (Verbert, Duval, Klerkx, Govaerts, & Santos, 2013).

Over the decades, discourse analysis on discussion forums has been carried out through various formats, network analyses, topical analyses, interactive explorers, knowledge extraction, etc. (Dave, Wattenberg, & Muller, 2004; Gretarsson et al., 2012; Indratto, Vassileva, & Gutwin, 2008; Lee, Kim, Cho, & Woo, 2013; Wei et al., 2010). With the rapid growth of free, open, and large user-based online discussion forums, it is essential, therefore, for education researchers to pay more attention to emerging technologies that facilitate learning in cyberspace. For instance, (Sande, 2010) investigated online tutoring forums for homework help, making observations on the participation patterns and the pedagogical quality of the content. (Hanrahan, Convertino, & Nelson, 2012; Posnett, Warburg, Devanbu, & Filkov, 2012) studied expertise modeling in such environment. Cohere (Shum, 2008) investigates semantic connections by identifying the link types to associate negative, positive, neutral interactions among online discourses. (Wise, Zhao, & Hausknecht, 2013) observed the listening behavior, which encapsulates different actions that learners take in relation to other’s posts (attending, reading etc.), to further describe the discussion engagement.

Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) is an unsupervised algorithm that uses “bag of words” approach to perform statistical topic modeling, which is a well-established method for uncovering hidden structures in large text corpora. There are several variations of LDA-based topic models successfully encapsulate large text semantics into topic words, such as online reviews, political opinions, microblog streams, email summaries etc. (Jo & Oh, 2011; Lan, Buntine, & Huidong, 2010; Liu et al., 2012; Wang, Agichtein, & Benzi, 2012). In this work, we present a novel Topical Facets Modeling method to capture online forum posts semantics.

3. TOPIC FACET MODEL

In order to automatically detect topics from conversational and relatively short amount of texts in each forum post, we propose a probabilistic topic model, Topic Facet Model, extending from LDA (Blei et al., 2003). A topic is a multinomial distribution of words that represents a concept from each forum post. A facet is a multinomial distribution of words that represents a more specific topic in the forum, for instance, extends (a java keyword) is one of the main facets in determining whether a program implemented inheritance concept in Java programming language or not. Thus, Topic Facet Model firstly adopts SLDA (Lan et al., 2010) in the topic model. Essentially, SLDA takes into account the position of each individual word of topic inference. It then forces all words in a sentence are generated from one topic. When a post is topic-specific, short-and-sweet, such as “how to write a for loop?”, SLDA is supposed to distinctively generate the corresponding topic word - loops. However, as we discussed earlier, an open discussion forums often mix with various complexities of posts. For instance, “Can an array of objects be iterated in enhanced for loop?” Given the sentence combines two main concepts, arrays and loops, SLDA will constrain only one topic word to be generated. In this case, the key of the question is about topic arrays (whether one can perform a function with array data structure), however, due to that there are more topic loops-related words represented, the SLDA will misinterpret it. This is where the facets come into play, to take into account specificity of a topic in the model. Following the same example, we can specify “array iteration” as a facet for topic loops. To explain Topic Facet Model algorithmically, Figure 1 shows the plate diagram.

The words generative process is explained following.

Figure 1. Topic Facet Model

1. For every pair of topic word \( t \) and facet \( f \), draw a word distribution \( \phi_t \sim \text{Dirichlet}(\beta) \)
2. For each document \( d \),
   a. Draw the document’s topic word distribution \( \pi_d \sim \text{Dirichlet}(\gamma) \)
   b. For each topic word \( t \), draw a facet distribution \( \theta_{tf} \sim \text{Dirichlet}(\alpha) \)
   c. For each sentence,
      • Choose a topic word \( j \sim \text{Multinomial}(\pi_t) \)
      • Given topic word \( j \), choose a facet \( k \sim \text{Multinomial}(\theta_{jk}) \)
      • Generate words \( w \sim \text{Multinomial}(\phi_{jk}) \)

Topic Facet Model exploits prior topic-keyword information by using asymmetric \( \beta \). For example, we expect “superclass, subclass, inheritance” are not probable in questions and answers on topic loops and similarly the words “for, dowhile, loops” is not synonymous to inheritance. This can be encoded into \( \beta \) such that elements of \( \beta \) corresponding to one topic-keyword have small values for the other topic-keywords. The latent variables \( \theta, \pi \) and \( \phi \) are inferred as Gibbs sampling as in original topic model (Blei et al., 2003). At each transition step of the Markov chain, the topic
keyword and facet of the \( i \)th sentence are chosen according to the conditional probability. The notations are described in Table 1.

\[
P(f_i = j, t_i = k | l_{i-1}, \ldots, l_1, \omega_i, \psi) \propto \frac{\gamma(T W_{k_j} + \beta_{jw})}{\gamma(T W_{k_j} + \beta_{jw} + m_{jw})} \frac{\gamma(BD_{d_j} + Y_j)}{\gamma(BD_{d_j} + Y_j + 1)}
\]

The approximate probability of keyword \( j \) in post \( d \) is

\[
\pi_{d_j} = \frac{C_{d_j}^{DP} + Y_j}{\sum_{j'=1}^{J} C_{d_j}^{DP} + Y_{j'}}
\]

The approximate probability of facet \( k \) for topic-keyword \( j \) in post \( d \) is

\[
\theta_{djk} = \frac{C_{djk}^{DFT} + \alpha_{jk}}{\sum_{k'=1}^{F} C_{djk}^{DFT} + \alpha_{k'}}
\]

The approximate probability of word \( w \) in topic-keyword facet \( \{ k, j \} \) is

\[
\phi_{jkw} = \frac{C_{jkw}^{FTW} + \beta_{jw}}{\sum_{w'=1}^{N} C_{jkw}^{FTW} + \beta_{jw'}}
\]

Table 1. Topic Facet Model notations

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
</table>
| \( D \) | number of posts, \( M \): number of sentences, \( N \): number of words, \( T \): number of topic-words, \( F \): number of facets, \( \omega \): word, \( t \): topic-word, \( f \): facet, \( \phi \): multinomial distribution over words, \( \Theta \): multinomial distribution over topic-words, \( \pi \): multinomial distribution over facets, \( \alpha \): Dirichlet prior vector for \( \Theta \), \( \beta_{(j)} \): Dirichlet prior vector for \( \phi \) of facet \( j \), \( \Upsilon_{(j)} \): Dirichlet prior vector for \( \pi \)

4. VISUAL ANALYTICS PROTOTYPE

In order to provide dynamic intelligent support for large-scale of online discussion forums, we build a Shiny web application by using statistical computing language, R. The web application (Figure 2) re-structures a discussion forum site into 3 parts: Filters, Analytics Visualizations and Forum Posts. They are represented in the following three UI panels from left to right:

1) Control panel (Left): data selection filters, such as post time range, post types (questions, accepted, non-accepted, all) etc. It also contains several semantic filters, such as topic models.

2) Analytics panel (Middle):
   - a summative forum post scatter plot (top): by configuring the x & y axes, it not only can provide the post density of entire discussion site over time, but also can indicate topic evolution.
   - topic facet model bubble chart (center): each bubble shows a facet associated with a topic, and the size of the bubble indicating the probability. For example, on selecting a post as Figure 2 shows, a set of corresponding topic words that represent Topic 2, where Inner facet particularly stands out to demonstrate the specificity.
   - learner-topic relation Sankey diagram (bottom): the user and topic relation diagram is purposefully designed to connect semantics and the communities, where can as well as be implemented in social network visualization.

3) Forum posts panel (Right): discussion posts for mutual referencing content and analytics.

![Stackoverflow Visualizer on java](https://example.com/stackoverflow.png)

Figure 2. Visual Analytics Prototype in Shiny, R
5. EVALUATION

5.1 Data Collection
We sampled one year (year 2013) of forum posts in topic Java from stackoverflow site through StackExchange API. The data pool was selected from the top 10 frequent tagged questions due to most the posts in this section contained at least one accepted answer. It will allow us to build a baseline on judging the topic facet model quality by comparing to non-accepted answers. There are total 16,739 posts, including 3,725 questions, 13,014 answers, with 3,718 accepted answers.

5.2 Automatic Forum Posts Topic Generation
Since the statistical language model only estimates the probability of words for natural languages, so we firstly parsed out all the program codes from forum posts. We then generate forum posts’ topics based on the Topic Facet Model. The topic generate process is reported in Figure 3. We found that accepted answers had significantly higher topic coverage than non-accepted answers (Mann-Whitney $U_{(M)}=606, Z=4.4362, p<.01$), where the topic coverage was computed by entropy. An example from the data set in the topic inheritance, where accepted answers topic facets covered a broader range of relevant concepts (privat, subclass, superclass, inherit, protect, public, class, object = inheritance), but none-accepted answers only covered a minimum subset of the topics (inherit, class, object, extend, subclass, super = inheritance). Note: we employed the state of the art Shannon entropy measure among several diversity measures considering a discussion post can cover a maximum unlimited number of topics.

Stack Overflow Forum Posts – Java

BagOfSentences.txt

topic-facet
java (0.0045)
class (0.012)
array (0.006)
....
Word List & Bag of Sentences preparation

Wordlist.txt

258 732 0 78
10

1

5.3 Topic-Facet Discovery
In order to capture the specificity from forum posts, we model content topics and the associated facets. To prove that topic modeling can be applied for conversational and relatively short sentences in forum posts, we compare the performance of TFM model with generic LDA. Table 2 demonstrates the discovered facets regarding to topic inheritance, where LDA was performed by MALLET toolkit with default $\alpha=30/N, \beta=0.01$. We can see that LDA seemed to generate a set of more sensible terms, but only a few of them are truly related to inheritance. On the other hand, TFM discovered the specificity from forum posts and uncover the topic related facets.

6. DISCUSSIONS AND LIMITATIONS
In this paper, we presented Topic Facet Model to extract topic specificity from conversational and relatively short texts from online programming discussion forums. We also presented a functional prototype of visual analytics for discourse-centric content. Our preliminary results demonstrated that TFM could be a promising approach to automatically generate discourse semantics for large-scale dynamic discussions. There are several limitations for current implementation. 1) We recognize that programming discussion forums are places for users to solve or to search for code solutions. It is also essential to include a code parser, to verify the solution quality as well as extract concepts from the codes. However, we argue that for a programming discussion forum, a natural language semantic parser and a program code parser can coexist and compliment each other. In the future study, we will consider both semantics parsers. 2) More exhaustive evaluation: The visual analytics is currently a proof of concept. More rigorous user study to examine the effects of auto-extracted semantics analytics on users is required. In addition, we did not discuss the topic facet model performance issues in this paper, however, when the topic numbers grow larger, which can be a serious factor for dynamic online support.

7. ACKNOWLEDGMENTS
The authors would like to thank Professor Denis Parra for helpful discussions. The research is supported by School of Computing, Informatics & Decision Systems Engineering.

8. REFERENCES

1 http://mallet.cs.umass.edu


