

Workflow-Based Assessment of Student Online Activities with Topic and Dialogue Role Classification

Jun Ma, JeonHyung Kang, Erin Shaw, Jihie Kim

Information Science Institute, University of Southern California
4676 Admiralty Way, Marina del Rey CA 90292, United States
{junma, jeonhyuk, shaw, jihie}@isi.edu

Abstract. The Pedagogical Assessment Workflow System (PAWS) is a new workflow-based pedagogical assessment framework that enables the efficient and robust integration of diverse datasets for the purposes of student assessment. The paper highlights two particular e-learning workflows supported by PAWS. The first workflow correlates student performance, as measured by project grades, with different dialogue roles, *information seeker* and *information provider*, that students take on in project-based discussion forums. The second workflow identifies the distribution of question topics within student discussions. Both workflows employ state of the art natural language processing techniques and machine learning algorithms for dialogue classification tasks. Workflow results were reviewed with a course instructor and feedback regarding the analysis and its fidelity are reported.

Keywords: Discourse analysis, workflow technology, discussion assessment.

1 Introduction

Online discussion forums are now an integral component of the virtual learning environments that are centrally supported by many colleges and universities, and have become an essential tool for student-student and student-instructor communication beyond the walls of the classroom. Course discussion forums contain rich information about student understanding of course concepts and assignments, and the resulting information provides invaluable feedback for instructors, allowing them to respond formatively to student concerns. However, even when instructors do participate in forums, they often do so question-by-question. In heavily used forums, patterns of participation can be impossible to discern, and patterns of discussion difficult to connect with course concepts. With better models for forum assessment it may be possible to better to identify misunderstanding and predict course performance.

As the use of online forums and other collaborative virtual learning technologies increase, the resulting heavier interactions introduce a considerable burden for teachers who wish to support their students' online activities. The Pedagogical Workflows project has developed a scalable e-learning framework to support efficient and robust integration of diverse datasets for the purposes of student assessment. The Pedagogical Assessment Workflow System (PAWS) employs the same computational workflow technologies that support scientific applications in the fields of seismology and astronomy [1]. PAWS facilitates the efficient processing and robust analysis of large amounts of data. A grid computing service acts as the backend of the system [2]. These existing workflow generation and execution approaches are applied to make

online assessment accessible to instructors. Workflow results are used to answer questions and provide formative feedback to instructors to facilitate “just in time” instructional adaptation to students learning and needs.

Recent work on integrating state of the art topic modeling and dialogue role classification techniques into PAWS is presented in this paper. The resulting classification results were correlated with other types of data, including questionnaire responses and project grades, through the workflows. Initial feedback on the resulting analysis was collected from a course instructor whose student discussions were fed on-demand into PAWS through a data collection service. The goal was that instructors would directly benefit from these new text tools.

2 Topic and Dialog Role Classification

The following sections describe the classification techniques used in PAWS.

2.1 SVM Classification Models for Online Discussion Threads

Support vector machine (SVM) is a widely used model in computer science and machine learning to perform classification tasks. PAWS uses SVM to classify both types of messages, i.e., *question* or *answer*, and types of users, i.e., *information seeker* or *information provider*, with respect to their dialogue roles in discussion forums. These classifications are important for the following reasons:

1. A student’s dialogue role indicates whether the student is asking for help or providing help to others. One cannot assume, for example, that every response provides an answer to a question; e.g., students with similar problems will sometimes join threads once initiated.
2. Knowing whether a piece of discussion text is a question or an answer (or neither) supports modeling of the types of discussions students engage in.

Analyzing individual messages with respect to their true *information seeking* or *information providing* roles is challenging. Standard surface-level grammatical forms are not enough to distinguish questions from answers. Surface-level features such as *wh* words such as what, where, when and how, or punctuation, such as question marks, are not sufficient. For example, some answers are commonly provided in a form of a question, e.g., “Have you checked the Nachos Manual section 4.3?”, and sometimes questions are posted to provide help rather than to seek it. So the same text can play different roles depending on context.

To train the SVM model, a labeled dataset that had been constructed by human annotators was used. Questions and answers within individual messages were marked. For user roles within a thread, the annotator marked the role of each participant as *information provider* or *information seeker*. The annotation scheme was developed over three years by multiple annotators (>6) until sufficient agreement on the data was reached. The annotators shared and compared their annotations while they were developing the scheme. The data used in this work was marked by two annotators using the final annotation scheme. Table 1 shows the Kappa values for inter-annotator agreement on a data subset that consisted of 30 discussion threads with 99 messages. Kappa values were computed with independent datasets. For all categories the

annotators show a high level of agreement (> 0.8). Then, a collection of feature templates was designed based on Kang et al. [3]. The features included word-based features such as uni-gram, bi-gram and tri-gram phrases, and discussion context features such as the position of current post in the thread. We also apply feature selection [3] to remove the noise and improve the performance.

Classifier	Precision	Recall	F-Score	Kappa
Question	0.88	0.88	0.88	0.93
Answer	0.83	0.80	0.83	0.96
Information Seeker/Provider	0.84	0.84	0.84	0.99

Table 1. Test Set Results on Question, Answer, and Information User Role

For this test, 240 discussion threads (904 messages) were randomly divided into two datasets: 180 discussion threads (634 messages) were used for training and 60 discussion threads (270 messages) were used for testing. Table 1 shows the model accuracy compared to the annotated target value. Results accuracy was almost 90%.

Role	Number initial and reply posts	SVM Classifier results for all discussion participants	SVM Classifier results for enrolled students only
Seeker	275	506	477
Provider	739	508	125

Table 2. Number of seeker/provider user roles in different settings.

The use of SVM classifier for student dialogue roles is obvious. In an initial implementation, a simple approach assigned the initial poster the role of seeker and all reply posters the role of provider. The approach was not accurate because students commonly seek information in the middle of a discussion thread. Table 2 shows the difference between the initial approach and the SVM approach. In the last column, we show the number of information seekers and providers for only those who received a course grade, which excluded the instructor and assistants. The results clearly show how much the instructor and assistants acted as information providers.

2.2 Topic Analysis on Student Online Discussion Text

Earlier interviews with instructors indicated that instructors were quite interested in topic-related discussion assessments [2], such as the topics of questions raised in the forum and their classification using topic categories from the course syllabus. As one of our objectives was to develop an approach that could be easily applied to different courses, supervised approaches requiring a large amount of labeled data were not appropriate. And because discussion datasets are noisy we needed a model that could capture semantic meanings behind the words rather than words themselves. Latent Dirichlet Allocation (LDA) [4] enables the capture of underlying semantic meaning without requiring large amounts labeled data, however, the original unsupervised LDA model was unsuitable because the topics learned by LDA are usually clusters of co-occurring terms that are not necessarily linked to real course topics. A semi-supervised model that could make use of course materials, such as syllabi and assignments was needed. The Labeled LDA model [5] was found to be appropriate.

The Stanford Part of Speech (POS) tagger was used first to extract nouns, since nouns in discussion sentences are the main indicators of the topics. Common words were then filtered out using a course-term dictionary that was semi-automatically

generated from the words in the assignment documents. Using Labeled LDA, each topic was profiled using a bag of words model, and then labels were assigned to discussion posts according to the topic bag of words. The labels act as a prior of topic distribution and thus affect the topics learned. For experiment and illustration, the Labeled LDA model was run using ten semesters of online discussion data and course materials. Fifteen topics were extracted. Table 3 shows five of the extracted course topics and their top N term lists.

Course topics	Most frequent words
Nachos Issue	function, call, line, class, type, code, nacho, thread, code, kernel,
Simulation	thread, custom, line, manager, clerk, number, switch, loop, problem,
Locks & Condition	lock, thread, condition, queue, wait, code, class, custom, variable, test,
Programming Issue	server, message, request, time, lock, system, error, code, array, char,
File System Call	file, page, swap, swap file, memory, bit, dirty, problem, size, swapfile

Table 3. Extracted course topics with their top N term list.

Although the Kappa value for agreement between two annotators was 0.96, the accuracies for the initial classifiers were low. Upon examination, several problems were found with the processing of student discussion data. First, the POS tagger did not generate correct results, especially because the system often failed to parse the noisy informal sentences that students wrote. It was also found that many irrelevant terms often misled the topic distribution process because LDA and Labeled LDA models regard each word/term in the document/thread equally when calculating the topic distribution of documents. The adoption of a domain ontology that is semi-automatically induced from a textbook glossary [6], to represent documents (discussion threads), might ameliorate these problems.

3 Assessment Workflows with Text Classification Components

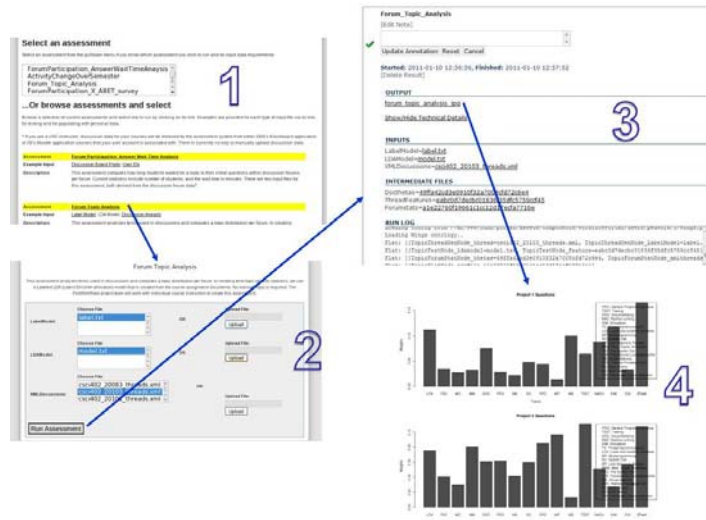


Figure 1. PAWS portal: The workflow user interface layer.

3.1 Computational Workflows for Student Learning Assessment

The workflow user interface layer, or PAWS portal, is shown in Figure 1. Steps 1-4 show how the system is used to run a sample assessment workflow and how the results are accessed. In Step 1, the user selects a student assessment workflow (template). In Step 2, the user specifies the resources (datasets) that will be bound to the workflow run instance. In Step 3, the workflow instance is submitted for remote execution [2]. In Step 4, the user views the results.

3.2 Relating Information Roles with Grades

A diagram of the *Role-Grade Analysis* workflow is shown in Figure 2. There are four *components* (data processing steps, shown in yellow) in the system.

1. DiscussionClassifier: Performs the SVM Classification on discussion text. The input resources include the discussion data, trained SVM model and n-gram feature model. The output is the classified text specifying student role per thread.
2. LinkGrades: Translates the instructor's XLS grade data into an internal format and links the IDs of graded students to the IDs of discussion participants.
3. RelateRolesWithGrades: Links dialogue roles and grades. The input datasets are ClassifiedRole (output of DiscussionClassifier) and XMLUserGrades (output of LinkGrades). The output is the RoleGradeTable, which specifies role and grade weights.
4. MultiBoxPlot: Presents the results as multiple box plots (Figure 3).

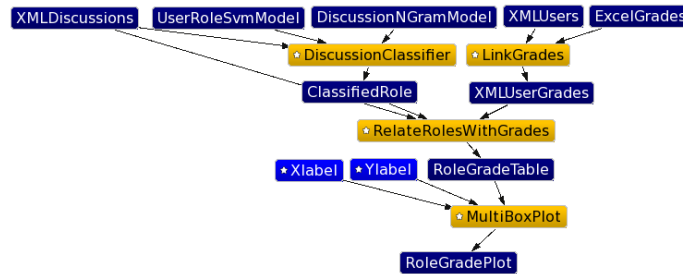


Figure 2. Diagram of the *Role-Grade Analysis* workflow.

The data flow is represented by the workflow diagram. Once the five input resources are selected by the instructor (or selected automatically by the system if there is only one matching dataset in the system, such as for the trained SVM models and n-gram feature models), a workflow instance is generated by Wings [1] and will be sent to the Pegasus [1] execution environment to run. Figure 3 shows a run result that used authentic data from an undergraduate computer science course.

The RelateRolesWithGrades component automatically rescales the grade level into five discrete levels and the BoxPlot component plots a box for each grade level. The box represents five values of the role weight distribution within the grade level: the starting point, the $\frac{1}{4}$ point, the median, the $\frac{3}{4}$ point and the end point. The level is a parameter of the workflow template so that instructors can change it for each run instance. The resulting graph shows a small trend: Students who perform better are more likely to be information seekers.

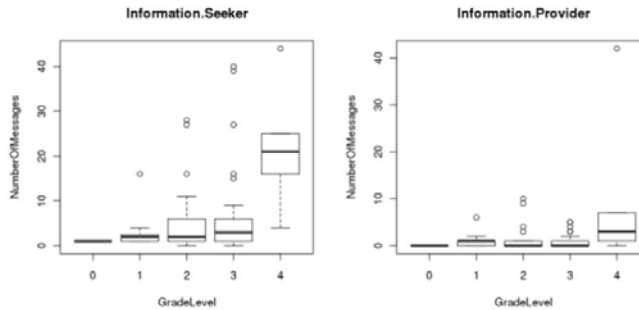


Figure 3. Result of Role-Grade Analysis workflow

3.3 Relating Questions to Topic Categories

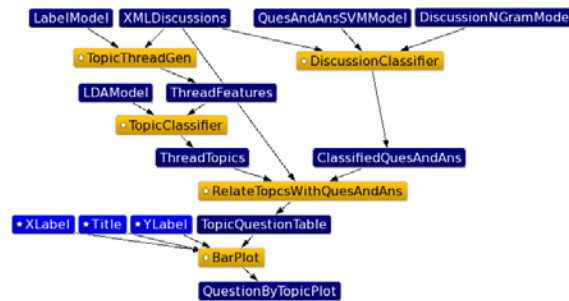


Figure 4. Diagram of QuestionByTopic workflow.

The second workflow integrates topic analysis and question-answer classification to identify the topic category that students ask the most questions about. This workflow directly addresses an assessment question that many instructors were interested in. The workflow consists of five components.

1. TopicThreadGen: Generates a feature vector for the Topic Classifier. The feature vector is the bag-of-words n-gram in the discussion text. Because a Labeled LDA Topic Model is used, it also assigns labels to each discussion thread via LabelModel.
2. DiscussionClassifier: This is the same component described in section 4.1 but performs Q/A rather than user role classification. The input SVM model is a trained Q/A SVM.
3. TopicClassifier: Determines the topic distribution given input discussion threads. The input is the trained LDA model.
4. RelateTopicsWithQuesAndAns: Links the discussion topics with discussion speech acts. Only questions raised are considered, so the output is the topic question table.
5. BarPlot: Presents the results as two bar graphs.

The resulting plots are shown in Figure 5. The top graph shows the number of questions raised in each topic category during the semester, while the bottom graph shows the number of distinct users raising questions in each topic category. The number of questions raised in each category is clearly different. During this semester, students asked questions about “programming assignment testing” and “memory

management”. This graph is of great importance to instructors for assessment purposes. The accuracy of the results will improve with the accuracy of the classifiers.

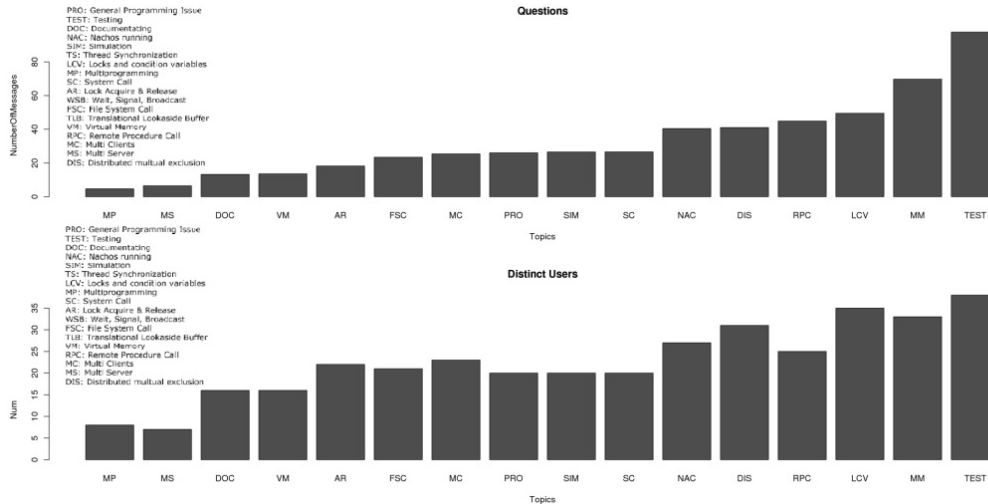


Figure 5. Results of the QuestionByTopic workflow.

4 Instructor Feedback

To collect feedback, the course instructor was given a description of the graphs and asked the following questions: 1) *Are the results understandable?*, 2) *How might you make use of the results?*, 3) *At what point during the course might it be helpful to have these results?*, and 4) *Do you have any suggestions for presenting the results?*

Regarding the role analysis, the instructor was able to understand the box plot and whiskers graphs but asked if real grades could be used instead of normalized grade levels. This would require that the instructor upload actual grades to the workflow instead of the absolute scores (0-40) used currently. The results confirmed for him that the best students were the most active and were not shy about asking questions when they had difficulties; and also that the providers understood these problems and had enough confidence to provide answers. He requested statistics about reading posts, venturing that “the top students read almost all, if not all, postings”. As far as making use of the results, he said that he could inform the class of these results, although he discounted the effect it might have.

Regarding the topic-based analysis, the instructor suggested that the results would be more useful if they a) reported why students posted questions, b) the topics were constrained to individual projects. The first comment indicates that a greater context will be necessary for assessment purposes. Regarding the second, although each project is assigned its own forum, the forums were aggregated for bettering machine learning results. The workflow can be modified to process results per project (i.e., per forum), but the results should be studied to ensure that no fidelity is lost.

5 Related Work

Researchers working on non-traditional, qualitative assessment of instructional discourse include [7]. As new assessments are developed and codified, they may be readily incorporated as components into the workflow system. Longitudinal studies of student performance [8] are also relevant and might be represented as workflows to electronically track student performance across courses.

6 Summary

This paper has demonstrated a new approach to processing and analyzing student information, especially data from online discussions, for the purpose of student assessment. Combined with traditional cognitive assessment methods such as assignment and exam grades, the workflow-based approach can be powerful tool for assessing impact of online learning. The approach utilizes NLP and machine learning techniques within the context of workflow, making both processing and analysis, both efficient and robust. Handling noisy student data and modeling subject topics were found to be very challenging tasks, primarily because existing NLP tools often failed to process discussion data correctly. To reduce variance, representing data using semi-automatically induced domain terms is currently being investigated. To increase accessibility of the assessment results, a weekly report of the workflow-processed results is being sent to the instructors.

Acknowledgement

The authors thank Senior Lecturer Dr. Michael Crowley for his support. This work is supported by a U.S. National Science Foundation CISE IIS award (#0917328).

References

1. Gil, Y., Ratnakar, V., Kim, J., Gonzales-Calero, P., Groth, P., Moody, J., Deelman, E. et al., WINGS: Intelligent Workflow-Based Design of Computational Experiments, IEEE Intelligent Systems. (2010).
2. Ma, J., Shaw, Erin., Kim, J. Computational Workflows for Assessing Student Learning. In *Proceedings of the 10th Int'l Conference on Intelligent Tutoring Systems (ITS2010)*.
3. Kang, J., Kim, J., Shaw, E. A Network Analysis of Student Groups in Threaded Discussions. In *Proceeding of Tenth International Conference on Intelligent Tutoring Systems (ITS2010)*.
4. Blei, D.M. and Ng, A.Y. and Jordan, M.I. Latent dirichlet allocation. *The Journal of Machine Learning Research*. Vol 3, 993—1022, 2003.
5. Ramage, D. and Hall, D. and Nallapati, R. and Manning, C.D. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*.
6. Feng, D., Kim, J., Shaw, E., and Hovy, E. 2006. Towards Modeling Threaded Discussions using Induced Ontology Knowledge. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI 2006)*. pp. 1289-1294.
7. McLaren, B. M., Scheuer, O., De Laat, M., Hever, R., De Groot, R., & Rose, C. P. (2007). Using Machine Learning Techniques to Analyze and Support Mediation of Student E-Discussions, , In *Proceedings of the 13th Int'l Conf. on Artificial Intelligence in Education*.
8. Reed-Rhoads (2008) C16 – Tools for Assessing Learning in Engineering. Presentation on Inventions and Impact 2: Building Excellence in Undergraduate STEM Education.