Towards automatic assessment of online discussions: Analyzing and modeling facets of user interactions

Abstract. With the increase in usage of online discussion boards for secondary education and other purposes, we need to be able to understand and model user interactions on these boards, to improve student learning, and user experiences. In this paper, we present both a corpus study analyzing how user's speech acts correlate with discussion characteristics such as length and following post type, and preliminary models for automatically classifying post states, where the possible states are *initiation*, *understanding*, *solving*, or *closing*, each corresponding to the actions posters take in the post. In the corpus study, we found a correlation between posts with simple *answer*-speech acts and short discussions, and a correlation between more complex *answer*-acts and longer discussions. Additionally, both preliminary state classifiers outperformed a simple random classifier.

1. Introduction

Internet discussion boards have become an essential tool for communication, in general, and in the context of higher education. Many colleges and universities now use course management systems that offer integrated discussion boards, furthering usage of discussion boards for higher education. With enrollments in online computer science and engineering courses increasing, and the increase in usage of online discussion boards in traditional classes, we wish to understand which student and instructor actions enable the best usage of online discussion forums for computer science students.

Additionally, the ability to automatically analyze question-answer style forums could benefit many help-based forums currently found on the internet, thus, we feel this type of work could be generalized to discussions outside of educational data, thus benefiting the discourse analysis and natural language processing communities.

In this paper, we focus first on analyzing annotated speech act tags already present in the data. We analyze how different types of answers to questions affect further discussion on the same or a similar topic. We compare answers that directly provide an answer to a students' question with answers that contain hints and elaborated answers, i.e. answers with a longer explanation. The preliminary results from these computer science students' discussions indicate that elaborated answers and hints may promote more participation from students and further discussions on related topics, in the form of longer threads. We also found preliminary evidence that these elaborated answers and hints may not only encourage new posts in the same thread, but promote more complex following posts. This work is currently under review[1].

We also investigated automatically classifying another already defined characteristic of discussion boards—state transitions between posts[2]. We attempt to classify four states—an "*initiation*" state, where students describe their problem, an "*understanding*" state, where others ask questions to try to understand what the initial problem is, a "*solving*" state, where participants try to develop an acceptable solution, and a "*closing*" state, where students generally indicate that an acceptable solution was found. Please see Section 3.1 of this paper, and Kang, et al for a more thorough discussion of these categories[2]. In this work, we investigated using supervised machine learning algorithms to classify these states. Both decision trees and hidden Markov models were investigated to attempt to model these classes.

2. Analyzing speech acts to model answers in online question-answer discussions

We investigate how different types of speech acts correlate with both length of discussion, and the types of post that follow that post. First, we discuss the data used for this corpus study, then we present the results of the corpus study.

2.1 Data

Our data comes from a computer science course at the University of Southern California, where students discuss homework problems and pose questions on the material and administrative topics. Each discussion thread represents a set of messages connected by reply-to relationships. Our corpus for this analysis consists of 385 discussion threads from two semesters of this course. These 385 threads come from two categories of forum posts—project-specific forums, and general-purpose administrative and lecture forums. 74 of the 385 threads do not contain an instructor presence.

For modeling Q&A dialogue and analyzing roles of different types of answers, we use Speech Acts[3], which can be grouped into the general categories of *question*, *answer*, *elaboration*, and *correction*. *Answer* sub-categories had a Kappa of 0.72, and *question* sub-categories had a Kappa of 0.94. Kappa is a score of agreement between annotators, which corrects for chance agreement. We split *answer* sub-categories into two categories: *answer-hint* (if the answer is given in the form of a hint), and *answer-direct*, for all other answer-types.

We investigate four types of posts—*single-direct*, where one single *answer-direct* type answer is given, *mult-direct*, where multiple *answer-direct* answers are given, *only-hint*, where only *answer-hint* are given, and *mult-ans-hint*, where multiple *answer-direct* answers are given, along with at least one *answer-hint*. Since this work focuses on the roles of these answer-types, we exclude threads that do not contain an *answer* speech act. 60 threads do not contain an *answer* act, and have an average length of 2.35 posts, while threads containing answer acts have an average length of 4.53 posts.

The distribution the analyzed threads can be found in Table 1. Figure 1 depicts the distribution of threads by their length.

	single-direct	*only-hint*	*mult-direct*	*mult-direct-hint*
Number of answer posts	483	31	365	2
% answer posts	52.39%	3.71%	43.66%	0.24%

Table 1: Distribution of Answer Types

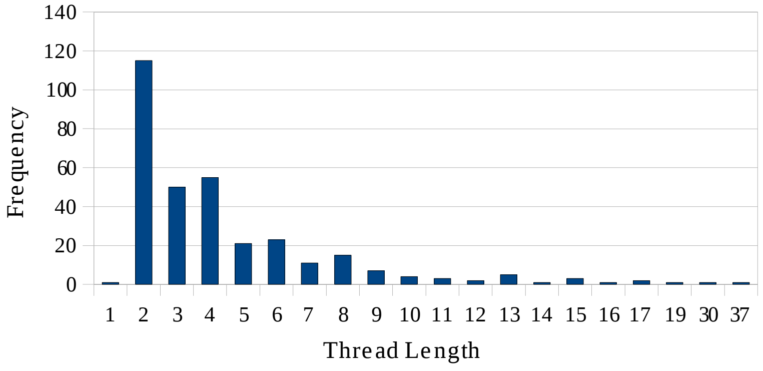

Figure 1: Thread-length Distribution

2.2 Answer-type's effects on thread properties

We first investigate how answer-type influences thread length. We divide the threads into three length-based categories: short, with 2 posts, medium, with 3 to 5 posts, and long, with 6 or more posts. We hypothesize that long threads contain more *mult-direct*, *only-hint*, and *mult-direct-hint* answer posts, while short threads contain more *single-direct* answer posts, i.e.

simple answers appear more frequently in short threads, and longer answers promote further participation.

Table 2 contains a count of these 'answer types'. Generally, there are more *single-direct* and *mult-direct* answers than other answer-types. Note that certain trends we expected to see appear in Table 2—the proportion of *single-direct* answers appears to decrease as threads get longer, while the proportion of *mult-direct* answers appears to increase. However, since long threads have more posts, we need to control for the number of answer posts per thread to make direct comparisons. Thus, Table 3 reports the average percentage of each answer-type found in each thread. We tested for significance using a two-tailed t-test, assuming unequal variance. Since we wished to compare the difference between long threads and short threads, the t-test was only performed between short threads and long threads, with medium threads not included in the analysis. Significant results (*$p<0.05$)* are denoted with an asterisk in Table 3, beside the percentage with the significantly larger mean.

The results of the t-tests confirm our hypothesis that the types of answers given in short threads and long threads are significantly different, in that there are more “short” (i.e. single) answers in short threads than there are in long threads.

Based on this result, we hypothesize that *mult-direct* answers encourage further student participation more than *single-direct* answers. We assess this hypothesis quantitatively in Section 2.3.

	Number of threads	*single-direct*	*only-hint*	*mult-direct*	*mult-direct-hint*	Number of *answer* posts
Short	115	93	4	46	1	144
Medium	129	142	8	118	0	268
Long	80	203	19	201	1	424

Table 2: Distribution of Answer Types by Thread Length

Avg. % of answer types per thread	*single-direct*	*only-hint*	*mult-direct*	*mult-direct-hint*
Short	66.52%*	3.48%	29.57%	0.43%
Medium	57.25%	2.87%	39.88%	0.00%
Long	51.20%	6.20%	42.18%*	0.42%

Table 3: Distribution of Answer Types by Thread Length,Controlling for Answers per Thread
** Denotes significantly greater at $p < 0.05$*

2.3 Effect of answer-type on next post

We also wish to investigate how answer-type impacts the next post, which we define as any post that directly replies to the current post. We split these next posts into five categories, dependent on the speech acts they contained—“question,” if the next post only contained *question* speech acts, “answer,” if the next post only contained *answer* speech acts, “both,” if it contained both *question* and *answer* speech acts, “other,” if it contained no *question* nor *answer* speech acts, and “final,” if no one replied to the current post. See Table 4 for the counts of these next posts, and Figure 2 for the percentages of each next-post type given each answer-type. For this analysis, we excluded 16 threads where reply-to relationships were not available in the annotation corpus.

Since we wish to compare the effects of *single-direct* answer-types to all other answer-types, we collapse *only-hint*, *mult-direct*, and *mult-direct-hint* into one category. We also directly compared *single-direct* and *mult-direct*, the most prominent answer types in our corpus. Table 4 also depicts the results of these significance tests, with ✢ denoting a difference between *single-direct* and all other answer types at *p<0.10*, and ✣ denoting a difference between *single-direct* and *mult-direct* with *p<0.10*. (Note that when comparing non-final next-posts, final next-posts were excluded from the analysis.)

Note that more *single-direct* answers are final posts than all other answer types. This supports our hypothesis that *single-direct* answers may not promote further contribution to the discussion. When following posts are present, question posts follow *single-direct* posts more than *mult-direct* posts. Thus, we further hypothesize that the short answer may not provide enough information to the information seeker.

	Question	Answer	Both	Other	Final	Total number next posts
single-direct (422 posts)	47✣	95	56	51	230 ✢	479
All other answer types (382 posts)	34	115	57	60	185	451
only-hint (31 posts)	4	5	2	7	14	32
mult-direct (349 posts)		110	55	52	170	417
mult-direct-hint (2 posts)	0	0	0	1	1	2

Table 4: Distribution of Next Post Types by Current Post's Answer Type
✢ Denotes single-direct vs all others significantly greater at p<0.10
✣ Denotes single-direct vs mult-direct significantly greater at p<0.10

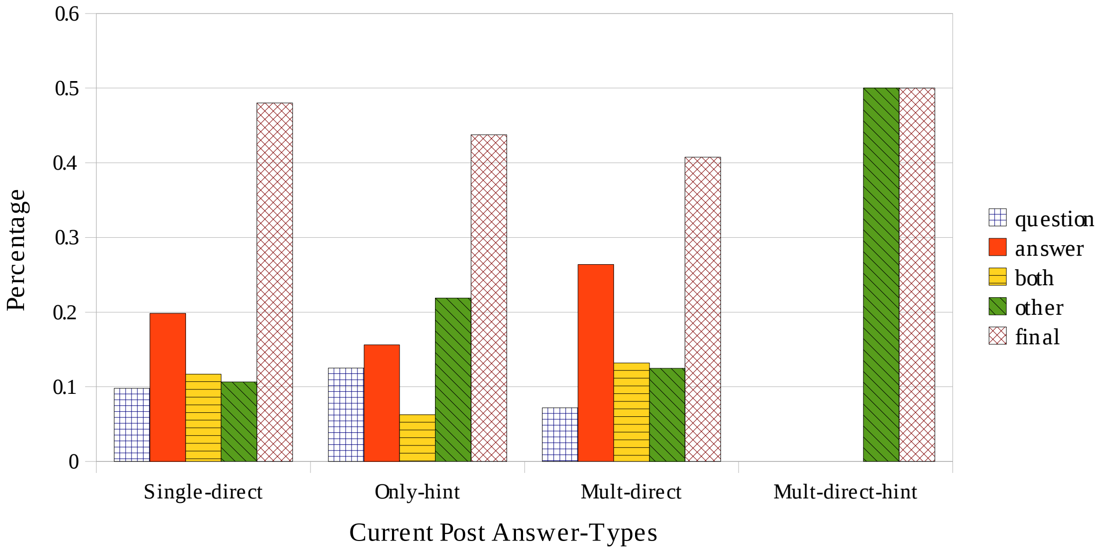

Figure 2: Percentages of Next Post Types, Split by Current Post's Answer Type

3. Modeling state transitions

Since the data used to model state transitions was a subset of the data used to analyze *answer* acts, we first describe this subset, then the methodology used to create the models, and then the results of training and testing the models.

3.1 Data

For these experiments, we use a subset of the data used for the speech act corpus study. This data consists of a subset of one semester's forums, which is annotated for both state transitions[2] and sink/source information[4]. The four states, *initiation*, *understanding*, *solving*, and *closing*, each correspond to a possible type of post typically found in our discussion threads—where users discuss their problem, attempt to understand what another person's problem actually is, solve the problem, and acknowledge that a solution worked, respectively. Annotator agreement for state transition had a final Kappa of 0.8405.

Sink/source information describes both the characteristics of the post, and the characteristics of the poster. There are four types of sink/source information—*hasSink*, *hasSource*, *isProvider*, and *isSeeker*. *hasSink* indicates if the post contains a request for information, while *hasSource* indicates if the post gives information. Note that these two categories are not mutually exclusive—one post can give and request information. *isProvider* and *isSeeker* are mutually exclusive and describe the poster's main intention—if the poster wants to provide information, or is actively seeking information. Inter-annotator agreement for *hasSink* had a Kappa of 0.9292, *hasSource* had a Kappa of 0.9595, and *isProvider/isSeeker* had a Kappa of 0.9898[4].

Note that the annotation manual used to annotate state transitions explicitly mentions if the poster is an information provider or an information seeker, in addition to describing other characteristics of these states. *Initiation* only occurs with information seekers. All other states can occur with information providers or seekers, however, both *understanding* and *solving* states begin with information providers. Since both *understanding* and *solving* phases can contain multiple iterations of question-answer discussion, posts answering or further questioning those initial *solving* or *understanding* posts are also labeled *solving* or *understanding*, respectively.

A total of 73 threads, containing 254 posts, were used to build a model for state transition. 151 of these posts were labeled *solving*, 93 were labeled *initiation*, 8 were labeled *closing*, and 2 were labeled *understanding*.

3.2 Method

First, we decided to initially use all sink/source information as features for this classification problem. Upon inspection of the annotation manual, it seemed that most definitions could be written as a combination of sink/source information. Additionally, an automatic classifier for sink/source information already exists, with a F-measures (a score that combines recall and precision) of 0.88 for *hasSink*, 0.83 for *hasSource*, and 0.84 for *isProvider/isSeeker*[4]. However, since this is preliminary work, we chose to use the gold-standard human classified sink/source information.

After looking at the data, we chose two standard supervised machine learning algorithms to try to model these states. We chose to investigate using decision trees, or tree, which would classify each post individually, and hidden Markov models, or HMM, which would classify each thread as a whole. Decision trees were chosen for multiple reasons. First, upon inspecting the annotation manual, it became apparent that our feature-space was highly partitionable if we used sink/source information as our features. Since decision trees iteratively partition the feature-space, this seemed to be a natural fit. Also, since decision trees produce output that's easily human-readable, they are commonly chosen as a first-pass algorithm, since it’s easy to see what features are important. We chose to investigate hidden Markov models since they would take into account characteristics of the thread as a whole. However, unlike decision trees, which trains at the post-level, a HMM would train at the thread level, so we have “more” training data for the decision trees, at the cost of the thread’s characteristics.

To test the decision tree and hidden Markov model classifiers, we compare these classifiers with a classifier that randomly assigns labels, with the same distribution as the training set. We refer to this baseline classifier as *rand*. For each classifier, we trained and tested on one 70/30 split of the data (where 70% of the threads were used for training, and 30% were used for testing), and we also performed 10-fold cross-validation. *k*-fold cross-validation is a process where the data is randomly partitioned into *k* complementary subsamples, then *k* different models are built, each one testing on its own k^{th} portion of the data, and training on the remaining data[5]. We then analyze the data based on Kappa, accuracy (i.e. percent correct overall), precision (i.e, correctly classified in a category over total number of that category classified), and recall (i.e, percentage correctly classified per category). To calculate these measures for the 10-fold cross-validation, we use a weighted average for precision, recall, and accuracy, with the weight controlling for the number of posts each model classifies. We approximate Kappa by combining the results from all 10 models into one “model” and then compute Kappa (i.e., we create one combined confusion matrix).

3.3 Results

We first present the results from the 70/30 split, and then we present the results from the 10-fold cross-validation.

See Table 5 for the confusion matrices for the 70/30 split for *rand*, tree, and HMM, respectively. (I = *initiation*, U = *understanding*, S = *solving*, and C = *closing*.) Note that the hidden Markov model is the only model to correctly assign the *understanding* category. Table 6 directly compares the precision and recall for each model for each category, while Table 7 compares the Kappa scores and accuracy.

	I	**U**	**S**	**C**
I	11	0	17	2
U	1	0	0	0
S	16	0	22	1
C	2	0	3	0

	I	**U**	**S**	**C**
I	29	0	1	0
U	0	0	1	0
S	4	0	35	0
C	0	0	2	3

	I	**U**	**S**	**C**
I	28	0	0	2
U	0	1	0	0
S	5	0	33	1
C	1	0	0	4

Table 5. Confusion Matrix for Testing Data for 70/30 Split for rand, tree and HMM models

	Precision				Recall			
Model	I	U	S	C	I	U	S	C
Rand	0.3667	0.0000	0.5238	0.0000	0.3667	0.0000	0.5641	0.0000
Tree	0.8788	0.0000	0.8974	1.0000	0.9667	0.0000	0.8974	0.6000
HMM	0.8235	1.0000	1.0000	0.5714	0.9334	1.0000	0.8462	0.8000

Table 6. Precision and Recall for rand, tree and HMM models

Model	Kappa	Accuracy
Rand	**	0.4400
Tree	0.8064	0.8933
HMM	0.7943	0.8667

Table 7. Kappa and Accuracy for rand, tree and HMM models
***Kappa cannot be calculated, since observed agreement less than chance agreement*

First, note that all models predict *initiation* and *solving* the best; also note that these are the two most prominent features in this data set. Also note that the HMM more accurately classified the two less prominent classes—*understanding* and *closing*—than any other model.

Also note that the Kappa for two human annotators was 0.8405. This can be considered an upper limit, or a gold standard for this task, and both Tree and HMM classifiers achieve a Kappa of approximately .8.

Tables 8, 9 and 10 show similar results for the 10-fold cross-validation experiments. For the confusion matrices, the results of all ten models were combined (since all of the testing sets are complementary, each post in the entire data set is represented exactly once.) This combined confusion matrix was then used to estimate Kappa and compute a weighted average of accuracy, found in Table 10. Table 9 uses a weighted average from each model, which is equivalent to computing precision and recall from the combined confusion matrix.

	I	**U**	**S**	**C**
I	31	1	56	5
U	0	0	2	0
S	58	1	88	4
C	3	0	5	0

	I	**U**	**S**	**C**
I	91	0	2	0
U	1	0	1	0
S	31	0	119	1
C	0	0	1	7

	I	**U**	**S**	**C**
I	93	0	0	0
U	1	1	0	0
S	40	1	110	0
C	3	0	0	5

Table 8. Combined Confusion Matrix for Testing Data for 10-fold cross-validation for rand, tree and HMM models

	Precision				**Recall**			
Model	**I**	**U**	**S**	**C**	**I**	**U**	**S**	**C**
Rand	0.3370	0.0000	0.5828	0.0000	0.3333	0.0000	0.5828	0.0000
Tree	0.7398	0.0000	0.9675	0.7778	0.9785	0.0000	0.7881	0.8750
HMM	0.6788	0.5000	1.0000	1.0000	1.0000	0.5000	0.7285	0.6250

Table 9. Precision and Recall for rand, tree and HMM models

Model	Kappa	Accuracy
Rand	**	0.4685
Tree	0.7271	0.8543
HMM	0.6746	0.8228

Table 10. Kappa and Accuracy for rand, tree and HMM models
***Kappa cannot be calculated, since observed agreement less than chance agreement*

Once again, note that HMM is the only model to correctly classify any *understanding* instances. Also note that the values in Tables 5, 6, and 7 are slightly higher than the values presented in Tables 8, 9 and 10. Recall that only one split of the data was analyzed for

Tables 5, 6, and 7, so only one subset of the data was tested on. For Tables 8-10, all data was used as a test set, thus reducing the chances of getting “lucky” and picking an “easy” test set.

4. Discussion and future work

We have presented preliminary work that both analyzes interactions between thread characteristics and speech acts, and preliminary models for classifying state transitions. In addition, these models used features we can classify with high accuracy.

For the interaction analysis, we hypothesized, and found, that the percentage of posts containing only a single answer decreased significantly between short threads and long threads, while the percentage of posts containing multiple answers increased significantly between short and long threads. We also provided preliminary evidence that there is a relationship between single answers and final posts, and qualitative evidence that answer-type does impact thread characteristics.

With respect to modeling state transitions, our preliminary models achieve Kappa scores within the range of our human annotators. However, these preliminary models still rely on human annotated features, and thus present a best-case scenario. Thus, we wish to investigate using automatically classified sink/source information as our features. In addition, this will also increase the number of annotated threads from 73 to 196 (we have more threads annotated for state transition than sink/source). These additional threads could help our classifiers be more accurate.

As noted, the HMM did slightly worse than trees, while we would expect the HMM to outperform a decision tree. However, the HMM requires more data to accurately train than the decision tree. Thus, we hypothesize that once more data is available for training, the HMM will eventually outperform the decision tree, since it models transitions. We also hypothesize that the HMM will be more robust to noisy training data than the decision tree, as in the case of using automatically classified sink/source information.

We would like to use this classifier to be able to, in the future, do large corpus studies with this data. If we can build a state transition model that can generalize across semesters, we could have a very large data set, which is acceptably accurate, without the cost of manual annotation. Kang, et. al. propose using state information to determine if student’s questions were successfully resolved, but a small dataset hinders that analysis[2]. Additionally, more data for speech acts is manually annotated than for state transitions. So, we could attempt to link speech act trends, presented in this paper, and state transition trends to analyze threads. Eventually, these large-scale corpus studies could result in pedagogical suggestions for instructors using online discussion boards in their classes.

5. Related work

Many previous studies have investigated learning in online discussions. Perkins and Murphy measured individual engagement and critical thinking processes in online discussions by identifying student clarification, exploration-support-assessment, inference and strategy[6]. Similarly, Gunawardena, Lowe, and Anderson examined learning and knowledge construction by identifying cognitive activities, arguments, resources, and changes in understanding[7]. These are all characteristics that we would like to combine with our answer-effect analysis. Jeong found that gender differences in communication style did not produce significant differences in response pattern types amongst students participating in an online debate[8].

Hew and Cheung analyzed students’ degree of participation in student-facilitated forums, and inferred a relationship between the number of (unique) participants and the size of an online discussion. They also inferred a relationship between students’ mental state, called habits of mind, and a greater number of participant postings[9]. It would be interesting to see if either of these results transferred to discussions where the instructor is present. Additionally, Hew and Cheung called for more analysis on factors that impact the duration of online discussions[9], which this paper has contributed to.

Also, our analysis only deals with active, posting participants. May, George, and Prévôt propose a framework for analyzing students' actions while using a forum that could record students who read, but do not post[10]. This could also provide an interesting method of analysis for how answer-types change the way students interact on a discussion board.

Mazzolini analyzed differences in types of instructor participation with respect to thread outcomes using frequency and placement of instructor posts to explain student satisfaction surveys, posting rates, and thread lengths[11]. This analysis of instructor participation could be complementary to our answer-type analysis.

Jeong proposed investigating state analysis, using a different set of states than ours, with asynchronous online discussion threads. However, Joeng does not attempt to automatically label these states, instead focusing on the corpus study, as we propose in our future work[12].

References

[1] Drummond, J., and Kim, J. (2010). Role of Elaborated Answers on Degrees of Student Participation in an Online Question-Answer. *Under review.*

[2] Kang, J.H., Kim, J., and Shaw, E. (2010). Modeling Successful versus Unsuccessful Threaded Discussions. *Workshop on Opportunities for intelligent and adaptive behavior in collaborative learning systems*, 13.

[3] Kim, J., Chem, G., Feng, D., Shaw, E., and Hovy, E. (2006). Mining and assessing discussions on the web through speech act analysis. *Proceedings of the Workshop on Web Content Mining with Human Language Technologies at the 5th International Semantic Web Conference*.

[4] Kang, J.H., and Kim, J. (2010). Analyzing Message Influence in Online Discussions with a Role-Based Information Network. *Internal project report.*

[5] Cross-validation (statistics). *Wikipedia, the Free Encyclopedia*. <http://en.wikipedia.org/wiki/Cross-validation_(statistics)#K-fold_cross-validation>.

[6] Perkins, C., and Murphy, E. (2006). Identifying and measuring individual engagement in critical thinking in online discussions: An exploratory case study. *Educational Technology & Society*, 9: 1, 298-307.

[7] Gunawardena, C.N., Lowe, C.A., and Anderson, T. (1997). Analysis of a global online debate and the development of an interaction model for examining the social construction of knowledge in computer conferencing. *Journal of Educational Computing Research*, 17: 4, 397-431.

[8] Jeong, A., and Davidson-Shivers, G (2006). The Effects of Gender Interaction Patterns on Student Participation in Computer-Supported Collaborative Argument. *Educational Technology Research and Development*, 54: 6, 543-568.

[9] Hew, K.F., and Cheung, W.S. (2010). Possible factors influencing Asian students' degree of participation in peer-facilitated online discussion forums: a case study. *Asia Pacific Journal of Education*, 30: 1, 85-104.

[10] May, M., George, S., and Prévôt, P. (2007). Tracking, Analyzing, and Visualizing Learners' Activities on Discussion Forums. *Proceedings of the sixth conference on IASTED International Conference Web-Based Education*, Vol. 2, 649-656.

[11] Mazzolini, M., and Maddison, S. (2003). Sage, guide, or ghost? The effect of instructor intervention on student participation in online discussion forums. *Computers and Education*, 40: 3, 237-253.

[12] Jeong, A. (2005). A guide to analyzing message-response sequences and group interaction patterns in computer-mediated communication. Distance Education, 26: 3, 367-383.