Modeling and Assessing Student Activities in On-Line Discussions

Jihie Kim Erin Shaw Donghui Feng Carole Beal Eduard Hovy

Information Sciences Institute
University of Southern California
Marina del Rey, CA, 90292
{jihie, shaw, donghui , cbeal}@isi.edu

Abstract

As web-enhanced courses become more successful, they put considerable burdens on instructors and teaching assistants. We present our work on developing software tools to support instructors by automatic assessment of pedagogical discussions. We are developing prototype measures of discussion quality that rely on the quantity of discussion contributions. We are also developing techniques for assessing discussion contributions automatically by mining discussion text. Using information retrieval and natural language processing techniques, our tools learn to detect the conversation focus of threaded discussions, classify topics of discussions, and estimate technical depth of contributions. The results from these assessment tools provide basis for the development of scaffolding and question answering techniques for pedagogical discourse.

Introduction

Web-enhanced courses and distance education courses are becoming increasingly popular. Such courses make class materials easily accessible to remote students, and increase the availability of instructors to students beyond the traditional classroom. Engagement in on-line discussions is an important part of student activities in distance education, and instructors often use it to measure each student's contribution to the class. However, as such courses become more successful, their enrollments increase, and the heavier on-line interaction places considerable burdens on instructors and teaching assistants. Thus, the ultimate success of web-based education is constrained by limited instructor time and availability. It is probably not feasible or pedagogically appropriate to automate completely the grading of on-line discussion contributions. However, if we can find a way to semi-automate some of the grading, then instructor time can be allocated more effectively to the particular students or discussion cases that truly require in-depth human monitoring and assessment.

We are developing prototype measures of discussion quality that rely on the quantity of discussion contributions. Most discussion board systems record the number of messages students post, which is a very crude indicator of participation. We may infer that a student is at least engaged with the class, relative to a student who never logs on to the discussion board at all. Number of posts can be significantly supplemented by including the number of responses that a post elicits from classmates and/or the TA or instructor. Posts that engage many responses might be particularly insightful, provocative, and thought provoking. Several such quantitative measures have been developed to assess on-line discussion activities (Kim and Beal 2006). Here we are validating the measures by applying them to two different courses with very different settings of discussion activities and relating them to the actual discussion grades and the instructor ratings. We focus on the number of posted messages, length of messages and number of responses that a post elicits from classmates and/or TA or instructor.

We are also developing techniques for assessing discussion contributions automatically by mining discussion text. Past approaches to mining information from discussion board text focused mainly on finding answers to questions (Feng et al, 2006b; Marom & Zukerman, 2005). Most of these techniques simply consider discussion data as text corpus. However, there are increasing needs for modeling discussion activities more explicitly. A discussion thread consists of a set of messages arranged in chronological order where people may express their ideas, elaborate arguments, and answer others' questions; many of these features in threaded discussions are unexplored by traditional IR techniques. Instructors want to review student discussions in order to understand what kinds of contributions are made by the students and whether they need any assistance or guidance (Painter et al., 2003). We may need to identify undesirable distractions, including contributions that are unrelated to the main focus. To support such assessment, we must be able to track the topics of discussion and determine if the contributions are focused and productive.

To support these capabilities, we have developed several techniques for modeling discussion threads. We consider on-line discussion a special case of human conversation. Each discussion thread contains a set of ordered messages from two or more people. The contents of a message, relations among the messages in a thread, and relations between each message and the thread to which it belongs are systematically analyzed. In particular, we model discussion threads as a sequence of speech acts and investigate dependencies among the messages using a set of relational dialogue rules (Feng et al., 2006a). This model supports our assessment of the most informative or important message in the sequence for the purpose of addressing the issue or question raised by the initial message (Feng et al., 2006b), and our analysis of discussion topic focus, including topics of individual messages and their relations to the topics of the discussion thread (Feng et al., 2006c).

The paper begins with a set of prototype measures of discussion quality that rely on the quantity of discussion contributions. We show relations between these measures and manual discussion assessment results. The following section presents several modeling approaches for threaded discussions: speech act classification, rhetoric analysis and topic identification. We conclude with directions for future research.

Validating Quantitative Measures with Discussion Grades and Instructor Ratings

The courses we have analyzed with quantitative measures are an undergraduate Psychology of Women course at the University of Massachusetts and a graduate-level Advanced Operating Systems at the University of Southern California. Both of them were held in 2003.

The psychology course included over 300 undergraduates. WebCT was used as a required course supplement to the in-class lectures. Students were assigned to virtual discussion groups of 10 students, yielding 30 groups. Discussion contributions were hand-graded by the instructor and the teaching assistants. Participation was optional but for those who participated, the discussion grades were used in computing the final course grade. Since discussions were initiated by the instructor who provided specific discussion topics, although the students could initiate some sub-threads, all of them were closely related to the original topic. The instructor and TA were monitoring the posts and participated in some of the group discussions. There were four discussion assignment sessions and we have analyzed one of them. Although the participation was optional about a half (131) of the students participated in the session that we have analyzed.

The computer science course had over 80 graduate students enrolled. Its on-line discussion forum was divided into 17 sub-forums following the 14 main themes of the operating systems course and several general issues such as course information, assignments, and suggestions for the course. However, the students could post any messages on any topics at any time. They could also start new threads on any of the themes. In fact most of the discussions were initiated by the students. Their participation was reflected in the class participation scores in combination with other class activities, consisting up to 10% of the final grade. Compared to the psychology course, the instructor made use of the student activities in the discussion forum in a rather informal way, assessing only whether a student's contribution was strong or weak.

Results from Student Discussions in the Psychology Course

For both courses we have used quantitative measures consisting of (a) total number of posted messages, (b) total length of all the messages posted, and (c) an estimation of the total number of messages elicited from the posts. In estimating the number of messages elicited by a post, we counted the number of the following messages in the same thread. Figure 1 shows the results from the psychology course. Since the discussion grades were available we could relate these three measures to the discussion grades.

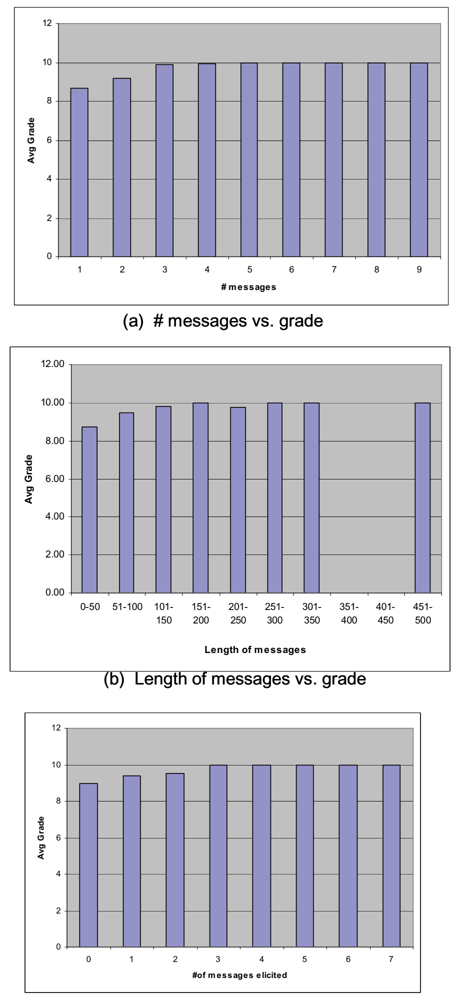


Figure 1. Degree of discussion participation vs. grade in the psychology course.

Figure 1 indicates that although most of the students received relatively good grades, the student who posted more messages, the students who posted longer messages and the students who elicited more messages received better grade. Table 1 shows the ranks of some of the students in three different groups: 5 students with highest ranks, 5 students with middle ranks, and 5 students with lowest ranks. As shown in the table, the top 5 students who participated more and elicited more messages received better (full 10) grades.

	A: # messages (rank)	B: Length of all the messages (rank)	C: # messages elicited (rank)	Average rank	Grade
S-high-1	6 (4)	312 (2)	7 (1)	2.33	10
S-high-2	8 (2)	267 (5)	7 (1)	2.67	10
S-high-3	9 (1)	277 (4)	5 (6)	3.67	10
S-high-4	5 (8)	285 (3)	5 (6)	5.67	10
S-high-5	5 (8)	213 (8)	4 (10)	8.67	10
S-mid-1	3 (38)	97 (54)	1 (66)	54.67	10
S-mid-2	2 (68)	97 (54)	2 (36)	54.67	10
S-mid-3	3 (38)	92 (60)	2 (36)	55.33	10
S-mid-4	2 (68)	90 (62)	2 (36)	55.33	9
S-mid-5	3 (38)	82 (66)	1 (66)	56.67	10
S-low-1	1 (111)	27 (126)	0 (106)	114.33	8
S-low-2	1 (111)	27 (126)	0 (106)	114.33	9
S-low-3	1 (111)	21 (128)	0 (106)	115.00	7
S-low-4	1 (111)	21 (128)	0 (106)	115.00	9
S-low-5	1 (111)	20 (130)	0 (106)	115.67	9

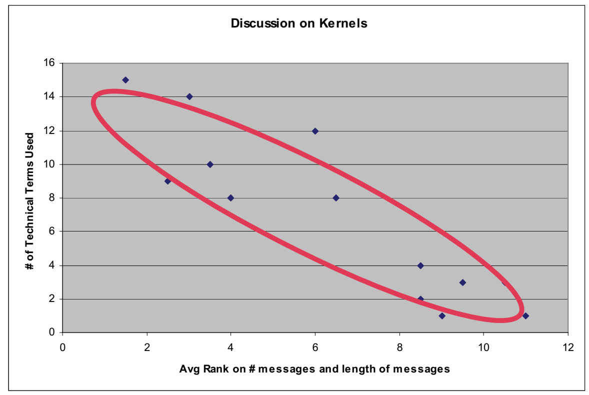
Table 1. Results from different groups of students in the psychology course.

Results from Student Discussions in the Computer Science Course

	A: Number of messages (rank)	B: Length of all the messages (rank)	C: #of messages elicited (rank)	D: #of threads initiated (rank)	E: #of different threads Participated (rank)	Average rank	Instructor's assessment
S-high-1	104 (1)	36726 (1)	507 (1)	16 (1)	37 (1)	1	strong
S-high-2	28 (3)	6790 (4)	96 (4)	4 (7)	18 (4)	4.4	strong
S-high-3	25 (4)	4285 (10)	92 (10)	8 (4)	23 (3)	5.2	strong
S-high-4	23 (6)	5174 (8)	120 (8)	5 (5)	16 (7)	5.8	strong
S-high-5	24 (5)	6708 (5)	95 (5)	3 (9)	17 (6)	6	relatively strong
S-mid-1	4 (29)	1331 (33)	21 (33)	4 (7)	3 (29)	24.6	not strong
S-mid-2	6 (22)	2182 (21)	45 (21)	0 (34)	2 (38)	25.2	not strong
S-mid-3	4 (29)	1143 (35)	24 (35)	1 (20)	4 (24)	26	not strong
S-mid-4	4 (29)	2602 (16)	23 (16)	0 (34)	3 (29)	26.2	not strong
S-mid-5	6 (22)	2100 (22)	13 (22)	0 (34)	4 (24)	26.8	not strong
S-low-1	2 (40)	275 (38)	0 (52)	0 (34)	2 (38)	43.8	not strong
S-low-2	1 (46)	345 (48)	5 (48)	0 (34)	0 (54)	44	not strong
S-low-3	1 (46)	178 (55)	3 (55)	0 (34)	1 (43)	44.2	not strong
S-low-4	1 (46)	325 (50)	1 (50)	1 (20)	0 (54)	44.4	not strong
S-low-5	1 (46)	579 (45)	1 (45)	0 (34)	0 (54)	46.2	not strong

Table 2. Results from different groups of students in the computer science course.

Table 2 shows the results from the computer science course. Since most of the discussion threads were initiated by the students and they could participate in any of the threads in any of the sub-forums, we have included two additional measures in this case: (d) number of threads initiated by the student and (e) number of different threads the student participated. If a student initiated more threads we may infer that he/she plays a leading role and introduces novel topics to the discussion than the students who elaborate or restate existing contributions. Also, if a student was involved in various discussions on different topics, we may infer that he/she has broader interests than a student who contributes to only small number of topics. The sixth column shows the average ranks based on these five measures. As shown in the table, the instructor agreed that in fact the top 5 students made strong contributions to the discussions with weaker contributions from others.



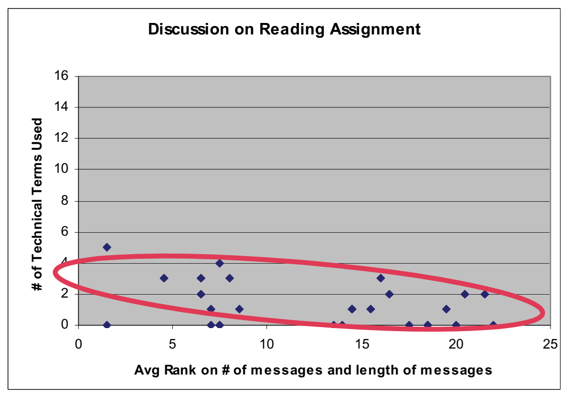


Figure 2. Usage of technical terms in different discussion threads.

Discussion contributions in the psychology course were very open in the sense that students could bring in various ideas and perspectives relevant to the given topic that are not necessarily taught in the class. However the discussions in the operating systems course were mainly about the concepts and techniques taught in the course and the instructor expected that technical discussions should refer to many of the technical terms that they have learned. In order to assess the kinds of contributions made by the students in the operating systems course, we have identified technical terms from the glossary in the operating systems text book. We have performed a simple stemming step to accommodate plural forms of the terms. Figure 2 shows our initial results from two popular discussion sub-forums: Kernels and Reading Assignment. The diagrams show the relations between the average ranks on the amount of contribution (i.e., the average rank on the number of messages and the total length of messages) versus the number of technical terms used. As shown in the figure in the technical discussion on Kernels, the students who contribute more (with higher ranks) tend to use more technical terms. However in the discussions on Reading Assignment, although a student contributes more and the rank with respect to the number of posts and length of the posts is high, the number of technical terms used can be very low, even down to zero.

Additional Findings from Quantitative Analysis

Unlike the discussions in the operating systems course, the instructor and the TA of the psychology course were closely monitoring discussion activities and participated in some of the group discussions. Their posts played various roles: providing an alternative perspective on the topic, supporting student presented ideas, elaborating student's answers, etc. The instructor and the TA participated in 17 group discussions (among 30 groups). The table below compares the average number of posts in the groups where the instructor and TA participated against the number without instructor/TA posts.

	Average # of Messages per Group
With Instructor / TA Participation	12.84
Without Instructor / TA Participation	15.19

Table 3. Effect on instructor/TA participation.

As shown in table 3, the groups with the instructor/TA participation had less number of posted messages. Contrary to our expectation, instructor involvement did not seem to increase student participation in the discussion. We are in the process of investigating the kinds of contributions that the instructor made and why the students posted fewer messages when there were the instructor/TA involvements.

Modeling Threaded Discussion

This section presents several approaches we have developed for modeling message threads in on-line student discussions. We exploit existing information retrieval and natural language processing techniques.

Speech Act Analysis

Conversation structures have received a lot of attention in the linguistic research community (Levinson, 1983). In order to integrate conversational features into our computational model, we must convert a qualitative analysis into quantitative scores. For conversation analysis, we adopted the theory of Speech Acts proposed by (Austin, 1962; Searle, 1969) and defined a set of speech acts (SAs) that relate every pair of messages in the corpus. Though a pair of messages may only be labeled with one speech act, a message can have multiple SAs with other messages.

We group speech acts by function into three categories, as shown in Figure 3. Messages may involve a request (REQ), provide information (INF), or fall into the category of interpersonal (INTP) relationship. Categories can be further divided into several single speech acts.

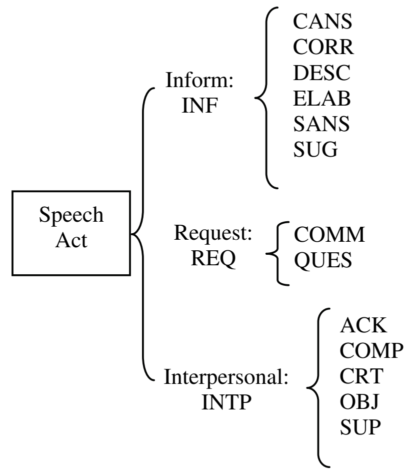


Figure 3. Categories of Message Speech Act.

Speech Act	Name	Description	Dir.
ACK	Acknowledge	Confirm or acknowledge	+
CANS	Complex Answer	Give answer requiring a full description of procedures, reasons, etc.	
COMM	Command	Command or announce	
COMP	Compliment	Praise an argument or suggestion	+
CORR	Correct	Correct a wrong answer or solution	–
CRT	Criticize	Criticize an argument	–
DESC	Describe	Describe a fact or situation	
ELAB	Elaborate	Elaborate on a previous argument or question	
OBJ	Object	Object to an argument or suggestion	–
QUES	Question	Ask question about a specific problem	
SANS	Simple Answer	Answer with a short phrase or few words (e.g. factoid, yes/no)	
SUG	Suggest	Give advice or suggest a solution	
SUP	Support	Support an argument or suggestion	+

Table 4. Types of message speech acts in corpus.

The SA set for our corpus is given in Table 4. A speech act may a represent a positive, negative or neutral response to a previous message depending on its attitude and recommendation. We classify each speech act as a direction as POSITIVE (+), NEGATIVE (−) or NEUTRAL, referred to as SA Direction, as shown in the right column of Table 4 (Feng et al., 2006b).

For evaluation we used an undergraduate level operating systems course corpus, which includes 3093 posts and 1236 threads. We first considered the distribution of the length of each thread (that is, how many posts were included in each thread), as shown in Table 5: 524 threads (over 40%) consist of only one post while most of the threads consist of from two to ten posts. Very few threads contain more than 10 posts. Compared to discussions in the graduate-level operating systems course, there seem to be fewer threads containing rich collaborative discussions.

Thread Length	Number of Threads
1	524
2	323
3	156
4	82
5	50
6	33
7	18
8	15
9	6
10	9
11	5
12	3
13	1
14	2
15	2
17	2
18	1
19	1
20	1
23	1
31	1

Table 5. Statistics of thread length in an undergraduate CS course.

Our corpus includes a total of 2173 Speech Acts. Table 6 shows the percentage of Speech Acts found in all posts of the annotated corpus.

We found that questions comprised the biggest portion of the corpus. This is consistent with the use of the board as a technical question and answer platform. Correspondingly, answers (CANS and SANS) and suggestions comprise 39.03% of total posts. The reason we consider suggestions together with answers is that for some of the questions, it is difficult to give an exact answer and in most cases, the replies are presented as suggestions. The ratio of complex answers to simple answers is 6.3. This matches our expectation that students ask lengthy context and procedural questions instead of simple factoid or Yes/No questions.

Code	Frequency	Percentage (%)
QUES	794	36.54
COMM	11	0.51
DESC	133	6.12
CANS	372	17.12
SANS	59	2.72
ELAB	149	6.86
CORR	25	1.15
OBJ	37	1.70
SUG	417	19.19
SUP	105	4.83
ACK	71	3.27

Table 6. Statistics of posted speech acts in archived discussions.

We also investigated the relations between two consecutive posts. As each post is classified as a Speech Act, the relations are represented by the consecutive relations between post speech acts. Table 7 gives the probabilities of transitions between all Speech Acts. To make it easier to understand, we add "START" and "END" states that refer to the start and the end of a thread discussion, respectively. Each represents the probability of going from the previous Speech Act (prev_SA in left column) to the next Speech Act (SA in top row). The information shows us how a discussion is conducted within a group of students. For example, there is a probability of 78.8% that any given discussion will start with a question (QUES), and a probability of 18.4% that it will start with a description of a situation (DESC).

Rhetorical Analysis

Rhetorical Structure Theory (RST) is a descriptive theory of the organization of natural text that grew out of studies of computational linguistics (Mann 1999). RST explains the coherence of text in terms of hierarchically-structured rhetorical relations that hold between two portions of text. We used an RST analysis of discussions to validate student reports that tutors helped scaffold discussions. SPADE (Sentence-Level Parsing of Discourse) is an RST discourse parser that purportedly achieves near-human levels of performance (defined as 90% accuracy) in the task of deriving sentence-level discourse trees (Soricut and Marcu 2003). We processed twenty-four online activities, constituting over one thousand message posts, during an on-line course in Distributed Learning at the British Open University (Shaw 2005). As shown in Tables 8 and 9, three relations generally stand out in tutor messages: attribution (the writer wants to make the owner of the text clear to the reader), elaboration (the writer wants to make it easier for the reader to understand), and enablement (whereby the writer wants to increase the potential ability of the reader). Other relations frequent in messages include *background,*

P(SA|prev_SA)	ACK	CANS	COMM	CORR	DESC	ELAB	OBJ	QUES	SANS	SUG	SUP	END
START	0	0	0.018	0	0.184	0	0	0.788	0	0.01	0	0
ACK	0.029	0	0	0	0.014	0.014	0	0.043	0	0	0.043	0.857
CANS	0.044	0.008	0	0.013	0.005	0.076	0.021	0.154	0	0.016	0.029	0.634
COMM	0.063	0	0	0	0	0.188	0	0.438	0	0.25	0.063	0
CORR	0.038	0	0	0.038	0.038	0	0	0.077	0	0	0.038	0.769
DESC	0.059	0.036	0	0.036	0.024	0.083	0.036	0.249	0	0.284	0.136	0.059
ELAB	0.052	0.091	0	0.006	0	0.143	0	0.117	0.013	0.071	0.013	0.494
OBJ	0	0.027	0	0.054	0.027	0	0.081	0.054	0	0.054	0.108	0.595
QUES	0.01	0.349	0	0.005	0.003	0.057	0.007	0.072	0.057	0.317	0.032	0.089
SANS	0.034	0	0	0	0	0	0.017	0.085	0	0.017	0	0.847
SUG	0.04	0.007	0	0.011	0.004	0.052	0.022	0.168	0.002	0.045	0.04	0.608
SUP	0.009	0.037	0	0	0.009	0	0.019	0.056	0	0.065	0.083	0.722

Table 7. Probabilities of speech act transitions.

cause, comparison, condition, contrast, and *explanation*.

Tutor-scaffolded activities (TGAs)									
role	#msg	attrib	bg	cause	cond	contr	elab	enbl	expl
tutor	172	4.9	1.2	0.2	0.8	0.6	10.6	1.6	0.1
student	492	3.0	0.6	0.1	0.4	0.3	5.5	0.7	0.0

Table 8. TGAs: Rhetorical relations as a percentage of messages posted.

Non-scaffolded activities (SGAs)									
role	#msg	attrib	bg	comp	cond	contr	elab	enbl	expl
tutor	26	26.5	1.5	3.9	0.0	0.0	51.2	4.2	1.92
student	401	6.8	1.6	0.3	0.8	1.1	10.7	1.6	0.2

Table 9. SGAs: Rhetorical relations as a percentage of messages posted.

Topic Classification

As a step towards modeling discussion threads, we want to identify topics discussed in threaded discussions and assess whether the topics shift or remain focused within the threads. Most machine learning approaches to topic classification use supervised learning techniques. They often require a set of manually labeled data, and the classifiers are trained with a selected learning algorithm, such as Naïve Bayes or SVM (Support Vector Machine). In most cases, manually labeling data is time consuming and expensive. Although some research proposes bootstrapping from limited data or explores the use of unlabeled data (e.g. Raskutti et al., 2002; Nigam et al., 2000), the need for a sufficient number of label examples remains a major bottleneck.

Furthermore, in an online discussion forum, the cost of labeling data may be bigger due to the following reasons. First, the total number of topics and the volume of messages are usually large, and the annotation of training examples for each topic is difficult and can easily become ad hoc, and this results in inconsistent annotations and noisy training examples. Second, messages in online forums are typically posted in chronological order, so it is not guaranteed that positive training examples for all topics exist in the corpus at the time of the annotation, and training for topics with sparse data is not possible. To overcome the lack of labeled data and reduce human annotation cost, we apply a Rocchio-style classifier to derive topic profile vectors through automatic ontology induction. In building topic profiles, instead of using a set of labeled data, we employ a coarse domain ontology that is automatically induced from a bible of the domain (i.e. the textbook). The details on ontology induction and the classification algorithm are described in (Feng et al, 2006c).

Since there are more rich discussions on technical topics in graduate-level courses, we have used discussions in the graduate level operating systems class for our analysis. No training data is required for the learning. However, in this particular course, the topic categories were given by the instructor in the syllabus so that students had to choose one of the categories when they initiated a thread. We use these manual annotations by the students as the gold standard for our analysis. There were 6 topic categories: 1:Communication Models, 2:Distributed Concurrency, 3:Naming and Binding, 4:Security, 5:File Systems, and 6:Case Studies. The categories correspond to one or more sub-trees in our ontology.

Thread Length	Number of Threads
1	5
2	17
3	7
4	5
5	2
6	7
8	1
9	2
10	2
12	1
16	1

Table 10. Thread length distribution.

The data set represents one semester of student discussions and comprises 206 messages and 50 threads. The average length of a thread is 4.12. Table 10 shows the number of threads by length in the corpus. Compared to the distributions in the undergraduate-level course (Figure 4), most of the threads (90%) consist of more than one post.

The distribution of the messages over topics according to our best classifier (Feng et al., 2006c) is shown in Figure 5. Most of the messages are classified into topics 1 and 5, with relatively fewer in topic 6.

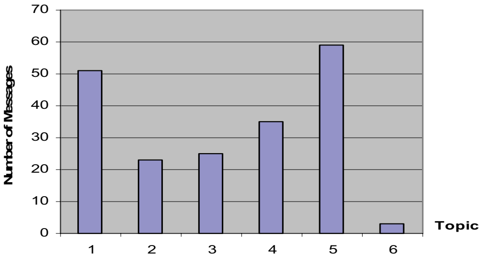


Figure 5. Statistics for topic-message distribution.

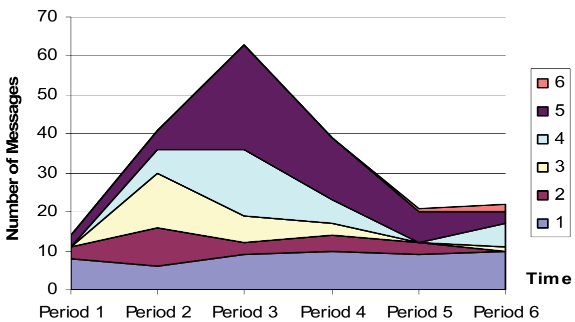


Figure 6. Temporal nature of topics

Figure 6 shows topic distribution changes over time. Each time period in the x axis represents a bi-week. The changes in the topic focus closely match the syllabus. Whenever the instructor starts a new topic, discussions on that topic will dominate on the discussion forum. The contributions include discussions on corresponding technical concepts, projects, and assignments.

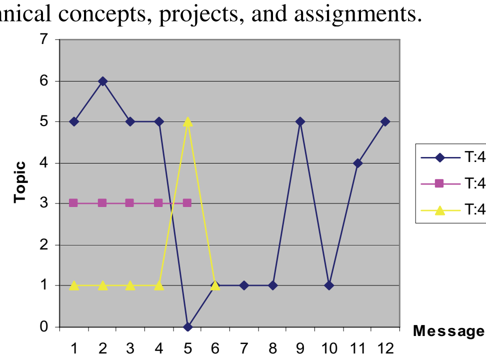


Figure 7. Topic shifting within threads.

When we investigated the details of topic shifts within each thread, we found more variances. Some discussions are very coherent while others have varying topics. Figure 7 shows topic shifts for three sample discussion threads. In Thread 4711, all the messages focus on the same topic, while Thread 4716 has only one message that leaves the main topic. Thread 4830 shows many changes in the topic.

Message 5 in Thread 4830 is classified as 'other' because it did not contain any terms defined in our ontology. These messages contain courtesy words or acknowledgements, such as 'Thank you' or 'It makes sense'. We are considering Speech Act methods to classify such messages more accurately.

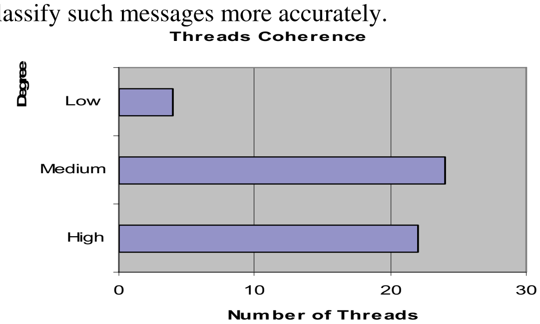


Figure 8. Thread coherence.

We computed the coherence measure of each thread by:

$$coherence = \frac{\max(\text{frequency of a topic})}{\#\text{ of messages}} \quad (10)$$

Using the coherence scores, we classified the threads into three categories: low, medium, and high, corresponding to the coherence interval [0, 0.4), [0.4, 0.8), and [0.8, 1]. As shown in Figure 8, most of the threads fall into high and medium categories. The results from these analyses can be used for information extraction or retrieval. For example, in retrieving answers to a question from a discussion corpus, we can remove irrelevant information and identify a coherent set of data sets that can answer the question.

Related Work

There have been other approaches to relating student learning activities to course materials. For example, Autotutor uses Latent Semantic Analysis (LSA) to evaluate similarity between student responses and the curriculum scripts (Graesser et al., 2001). LSA has been also used in grading student essays (Landauer 2002). Although the course discussions we have looked at are less structured, similar measures can be adopted in assessing technical quality and may be used in combination of other quantitative measures we are using.

There have been various approaches to assessing collaborative activities. For example, patterns of collaborative interactions in math problem solving have been analyzed by (Cakir et al., 2005). Various approaches of computer supported collaborative argumentation have been discussed (Shum 2000). Machine learning techniques have been applied to train software to recognize when students have trouble sharing knowledge in collaborative interactions (Soller and Lesgold, 2003). Our assessment techniques are broadly applicable in assessing various discussion activities and we believe that integrating our techniques with these capabilities may result in improved assessment of the kinds of contributions made by the students and predicting whether a teacher's involvement is needed or not.

Summary and Future Work

We are developing software tools to support instructors by semi automatic grading of discussions based on quantitative measures of discussion quality. We have developed several quantitative measures that rely on quality of discussion activities. The results from two courses show that the students who participate more and elicit more messages tend to receive better grades or ratings. Analysis of technical term usages in technical and non-technical discussions indicates that frequency of technical terms can supplement other quantitative measures by providing hints about the type of contributions students make.

Speech act classification results show that many threads in undergraduate discussions consist of only 1 or 2 messages and students do not fully exploit collaborative problem solving environment. We also have identified several relations that tutors use in greater numbers than do students as a means to scaffold discussions.

Fine-grained analysis of discussion activities may help us identify less productive and unfocused discussions where scaffolding is needed. In addition, extensive analysis of student discussion activities and discussion threads can support question answering by extracting useful information from the discussion corpus.

References

Cakir, M., Xhafa, F., Zhou, N., and Stahl, G. 2005. Thread-based analysis of patterns of collaborative interaction in chat, In *AIED-2005.*

Feng, D., Shaw, E., Kim, J., and Hovy, E.H. 2006a. An Intelligent Discussion-Bot for Answering Student Queries in Threaded Discussions. In *Proceedings of IUI 2006*.

Feng, D., Shaw, E., Kim, J., and Hovy, E.H. 2006b. Learning to Detect Conversation Focus of Threaded Discussions. In *Proceedings of HLT-NAACL 2006*.

Feng, D., Kim, J., Shaw, E., and Hovy, E.H. 2006c. Towards Modeling Threaded Discussions through Ontology-based Analysis. In *Proceedings of AAAI-2006*.

Fiscus, J. and Wheatley, B. 2004. Overview of the TDT 2004 Evaluation and Results. In *TDT-2004.*

Graesser, A.C., Person, N., Harter, D., & TRG 2001. Teaching tactics and dialog in AutoTutor. International Journal of Artificial Intelligence in Education, 12, 257-279.

Isbister, K. Nakanishi, H., Ishida, T., Nass, C. 2000. Helper agent: Designing an assistant for human-human interaction in a virtual meeting space. *Proceeding of CHI'2000*.

Joachims, T. 1997. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. In *ICML-1997.*

Kim J. and Beal C., 2006. Turning quantity into quality: Supporting automatic assessment of on-line discussion contributions, *AERA (American Educational Research Association) Annual meeting.*

Landauer, T. K. 2002 Applications of Latent Semantic Analysis. 24th Annual Meeting of the Cognitive Science Society, August 9th 2002.

Levinson, S. 1983. *Pragmatics*. Cambridge Univ. Press.

Liu, B. Li, X., Lee, W-S, and Yu, P.S. 2004. Text Classification by Labeling Words. In *AAAI-2004*.

Mann, W. (1999) An Introduction to Rhetorical Structure Theory (RST).

Marom, Y. and Zukerman, I. 2005. Corpus-based Generation of Easy Help-desk Responses. *Technical Report,* School of Computer Science and Software Engineering, Monash University.

Painter, C., Coffin, C. & Hewings, A. 2003, Impacts of Directed Tutorial Activities in Computer Conferencing: A Case Study. Distance Education, Vol. 24, No. 2.

Raskutti, B., Kowalczyk, A., and Ferra, H. 2002. Combining Clustering and Co-training to Enhance Text Classification Using Unlabelled Data. In *Proceedings of SIGKDD.*

Rocchio, J.J. 1971. Relevance Feedback in Information Retrieval. In G. Salton, *The SMART Retrieval System: Experiments in Automatic Doc. Processing*, pp. 313-323.

Shaw, E., 2005 .Assessing and Scaffolding Collaborative Learning in Online Discussions, *Proceedings of AIEd-2005*.

Shum, B. S. 2000. Workshop report: computer supported collaborative argumentation for learning communities, *SIGWEB Newsl*. 2000., 27-30.

Soller, A., & Lesgold, A. 2003. A Computational Approach to Analyzing Online Knowledge Sharing Interaction. *Proceedings of AI in Education*.

Searle, J. 1969. *Speech Acts*. Cambridge: Cambridge Univ. Press.

Soricut, R. and Marcu, D. (2003). Sentence Level Discourse Parsing using Syntactic and Lexical Information. Proceedings of HLT/NAACL 2003.