

Turning quantity into quality: Supporting automatic assessment of on-line discussion contributions

1. 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 work, 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 relies on the quantity of discussion contributions. For example, the number of posted comments is a very crude first indicator. We may infer that the 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 [2]. 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 also have explored an additional measure that takes into account the contribution content, i.e., the technical terms that students use during the discussions in a course.

The results from the two courses show that the students who participate more and elicit more messages tend to receive better grades or instructor ratings. Also our analysis of technical term usages indicates that frequency of technical terms provides hint on whether the content is technical or not and can supplement other quantitative measures. That is, although a student might score high on the number of posts, if the posts were non-technical such as asking for clarification of assignments or raising administrative issues, the score on the number of technical terms can be very low.

2. Validating quantitative measures with discussion grades and instructor ratings

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

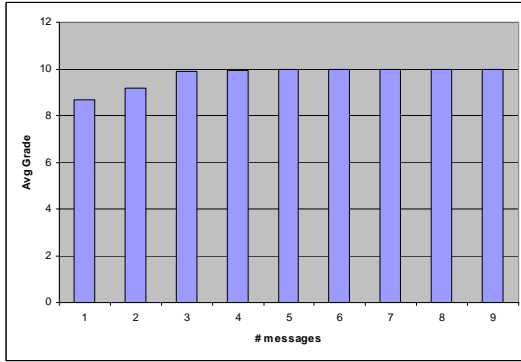
The psychology course included over 300 undergraduates. WebCT was used as a required course supplement to the in-class lectures. Students were assigned to virtual discussion groups of 10 students, yielding 30 groups. Discussion contributions were hand-graded by the instructor and the teaching assistants. 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.

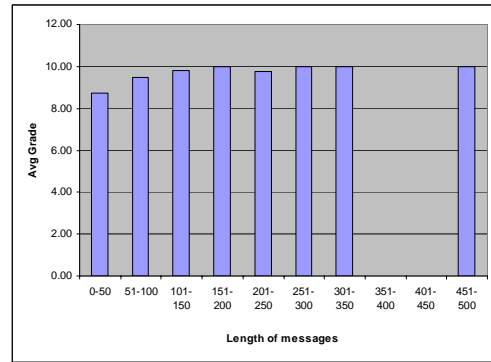
2.1 Results from student discussions in the psychology course

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.

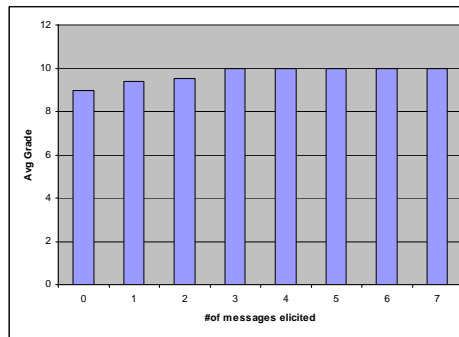
Figure 1 indicates that although most of the students received relatively good grades, the student who posted more messages, the students who posted longer messages and the students who elicited more messages received better grade. Table 1 shows the ranks of some of the students in three different groups: 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 vs. grade



(b) Length of messages vs. grade



(c) # messages elicited vs. grade

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

	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

2.2 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 and the contributions from others were less strong.

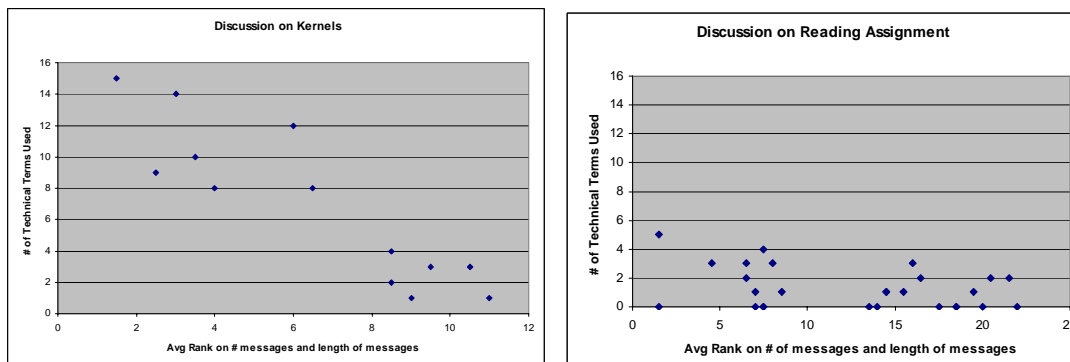


Figure1: Usage of technical terms in different discussion threads.

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.

3. 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 Number 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.

4. Related Work

There have been approaches to in relating student learning activities to course materials. For example, Auto-tutor uses Latent Semantic Analysis (LSA) to evaluate similarity between student responses and the curriculum scripts [2]. LSA has been also used in grading student essays [3]. 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 assess collaboration activities. For example, machine learning techniques have been applied to train software to recognize when the students have trouble in sharing knowledge through collaborative interactions [5]. Also a discourse parser was employed to assess the types of contributions made by the discussion participants including the instructor [4]. Our quantitative measures are broadly applicable in assessing various discussion activities and we believe that integrating our measures with

these capabilities may result in improved assessment of the kinds of contributions made by the students and predicting whether teacher's involvement is needed or not.

5. Summary and Conclusion

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 measures for assessing discussion activities and validated the measures by applying them to two different courses with very different settings of discussion activities. In particular we used the number of posted messages, length of messages and number of responses elicited and related them to the actual discussion grades and the instructor ratings. The results from the two courses show that the students who participate more and elicit more messages tend to receive better grades or ratings. Also additional 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 on the type of contributions students made.

References

- [1] 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.
- [2] Kim, J., Beal, C., and Maqbool, Z., (2005). Developing Teaching Aids for Distance Education, *Proceedings of AIED-2005*.
- [3] Applications of Latent Semantic Analysis Landauer, T. K. (2002). 24th Annual Meeting of the Cognitive Science Society, August 9th 2002.
- [4] Shaw, E. (2005) .Assessing and Scaffolding Collaborative Learning in Online Discussions, *Proceedings of AIED-2005*.
- [5] Soller, A., and Lesgold, A., Computational Approach to Analyzing (2003). Online Knowledge Sharing Interaction, *Proceedings of AIED-2003*.