Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Topic-Level Social Network and Language Correlation in Course Discussion Forums

Ezekiel Adriel LAGMAY[a*] & Maria Mercedes RODRIGO[a]
[a]*Ateneo Laboratory for the Learning Sciences, Ateneo de Manila University, The Philippines*
*ezekiel.lagmay@obf.ateneo.edu

Abstract: In this paper, we characterize student contributions in online discussion forums and examine the relationship between these characteristics and students' peer-to-peer relationships. We use Coh-Metrix Language and Discourse Analysis metrics to predict the Weighted Degree and Closeness Centrality of student contributions. We performed the analysis on all students in the population under study and on a subset consisting of active students only. We found that students who have more direct connections with other students tend to have abstract word choices whereas active students also tend to be more expository and informational, have shallow ideas, and use simpler construction. We also found that all students who can easily share their thoughts to the entire class tend to have posts that are more informational, with deep and connected thoughts and ideas. Active students who belong in this group further exhibit simpler construction and more abstract word choices.

Keywords: Online Course Discussion Forums, Social Network Analysis, Language and Discourse Analysis, Linear Mixed-Effects Models

1. Introduction

In this study, we characterize student contributions in online discussion forums and examine the relationship between these characteristics and students' peer-to-peer relationships. We use Coh-Metrix Language and Discourse Analysis metrics to predict the Weighted Degree and Closeness Centrality of a student's participation in a course discussion forum (Dowell et al., 2015). We use a Linear Mixed-Effects Modelling approach that also takes into consideration the differing nature of each course discussion forum topic within that same course. Our research questions for this study are as follows:

1. For each Social Network Analysis (SNA) metric (Weighted Degree or Closeness Centrality) and learner group (all students or active students only), which of the characteristics of language best predict that SNA metric of the student in a course discussion forum topic?
2. For each SNA metric (Weighted Degree or Closeness Centrality) and learner group (all students or active students only), does its correlation with the characteristics of language remain constant regardless of the method of measuring language and discourse features (percentile or z-score)?
3. What insights can we derive from these metrics about the quality of students' contributions and their relationships with their peers?

This study is based on a previous paper by Dowell et al. (2015). In this work, the proponents first extracted language and discourse scores from each student's course discussion forum post contents using Coh-Metrix and correlated these with that student's SNA metrics Degree Centrality, Closeness Centrality, and Betweenness Centrality. The Coh-Metrix scores used as independent variables include Narrativity, Deep Cohesion, Referential Cohesion, Syntactic Simplicity, and Word Concreteness (Graesser, McNamara, & Kulikowich, 2011; Graesser & McNamara, 2011; Kintsch, 1998; Snow, 2002). Dowell et al. (2015) generated two datasets: one containing all the students in the course and the other containing only the active students (those who have made 4 or more posts) (De Laat, Lally, Lipponen, & Simons, 2007; Gillani, Yasseri, Eynon, & Hjorth, 2014). A Linear Mixed-Effects Modelling approach was used with the Coh-Metrix scores as the independent variables, the SNA metrics as dependent variables, and learner (User ID) and word count as random effects (or data groupings) (Dowell et al., 2015; Knowles, 2013). The results showed that "more narrative style discourse with less overlap between words and ideas, simpler syntactic structures and abstract words"

positively increases the significance and centrality of a student in the social network (Dowell et al., 2015).

This study aimed to replicate the methodology used in the paper, but with the following changes:

- Usage of Topic ID as one of the random effects, and as well basing the definition of active students as those who participated in all discussion forum topics in a course, rather than basing it merely from the number of posts made by the student.
- Use Weighted Degree instead of Degree Centrality as one of the SNA metrics to be correlated.
- Experiment with the usage of either the percentile or z-score methods of language and discourse metrics to determine if their correlation with the SNA metrics still hold for both scoring methods.

> *The study made use of the discussion forum of Course A from a university in Metro Manila, Philippines which ran from August 26, 2021 to December 18, 2021. It had 20 enrolled students who made a total of 206 discussion forum posts, each grouped into seven distinct graded topics. 10 of the students have made at least one post in all seven discussion forum topics; these students are hence classified as "Active Students". Prior to the analysis, the course discussion forum dataset of Course A is split according to topic. Thus, each topic has a total of five datasets – one for post contents (which are then analyzed using Coh-Metrix), two for social network graphs (All Students and Active Students; these are then used for extracting the Weighted Degree and Closeness Centrality metrics via Gephi 0.9.3), and two for discussion topic participants (All Students and Active Students). Then prior to the analysis proper, the SNA and Coh-Metrix results for all topics are joined together into two datasets, one containing all students and another containing the active students only. In total, 6 models were generated, one for each possible Weighted Degree/Closeness Centrality-All Students/Active Students-Percentile/Z-Score combination.*

2. Results and Discussion

Weighted degree. For the All Students models, Narrativity and Word Concreteness significantly had a positive correlation and negative correlation with Weighted Degree, respectively. On the other hand, for the Active Students models, Deep Cohesion, Word Concreteness, and Narrativity had a negative correlation, while Syntactic Simplicity had a positive correlation. This could mean that students who had more direct connections with their fellow classmates in discussion forums in general created posts that had more abstract words. Students in the Active Students subgroup tended to be more expository and informational in their posts, had shallow ideas, and used simpler construction. Students in the All-Students subgroup who tended to be more narrative were also more likely to have more direct connections with others.

Closeness centrality. For all students, Narrativity had a negative correlation while Deep Cohesion and Referential Cohesion both had a positive correlation with Closeness Centrality. On the other hand, for active students, the correlation rules were the same, with the addition of the negative correlation with Word Concreteness and positive correlation with Syntactic Simplicity. This meant that students who could quickly disseminate responses to others tended to create posts that were more informational, with deep and connected thoughts, ideas, and words, while active students had the additional feature of using more abstract words and simpler construction.

Percentile vs. Z-Score. Narrativity, Referential Cohesion, and Deep Cohesion still held the same correlation rules whether percentile or z-score for both Weighted Degree and Closeness Centrality for the All Students models, and Deep Cohesion and Syntactic Simplicity for Active Students models. Hence, only Deep Cohesion still held the same correlation rules regardless whether percentile or z-score is used, regardless of learner group or SNA metric. The percentile-based models also had a higher percentage of the contribution of the independent variables to the predictable variance than the z-score-based models (Dowell et al., 2015).

3. Conclusion and Further Studies

In summary, some language and discourse characteristics had the same nature of correlation for all students in general and active students in particular, on a per SNA metric basis. However, determining the set of language and discourse metrics that could best predict both Weighted Degree and Closeness Centrality is only possible for the active students. Given the small sample size and the limitation of this study to a single class, future studies may want to determine if this finding continues to hold true for larger populations and across different subject areas.

This study also shows that, with the exception of Deep Cohesion, the nature and significance of the correlation between the SNA metrics and language and discourse features would differ in general depending on whether percentile or z-score is used as basis. The significance tests showed that, in this instance, models using the percentile-based scoring system are more predictive than those using z-score-based scores as they have a higher percentage of the contribution of the independent variables to the predictable variance.

Finally, we found that students who have more direct connections with other students tend to have abstract word choices whereas active students also tend to be more expository and informational, have shallow ideas, and use simpler construction. We also found that all students who can easily share their thoughts to the entire class tend to have posts that are more informational, with deep and connected thoughts and ideas. Active students who belong in this group further exhibit simpler construction and more abstract word choices.

Acknowledgements

The authors would like to thank the Ateneo Laboratory for the Learning Sciences, Ateneo Research Institute for Science and Engineering (ARISE), and Accenture for the funding and support needed for this research. The authors would also like to thank Deni Jaramillo and Miguel Saavedra for their assistance in setting up the necessary servers for the collection of Canvas data. The authors would also like to thank Natalie Newton and Danielle McNamara for providing the Coh-Metrix desktop application for the purposes of this study.

References

De Laat, M., Lally, V., Lipponen, L., & Simons, R.-J. (2007). Investigating Patterns of Interaction in Networked Learning and Computer-Supported Collaborative Learning: A Role for Social Network Analysis. *International Journal of Computer-Supported Collaborative Learning*, *2*(1), 87-103.

Dowell, N. M., Graesser, A. C., Hennis, T. A., Skrypnyk, O., Dawson, S., de Vries, P., … Kovanović, V. (2015). Modeling Learners' Social Centrality and Performance Through Language and Discourse. In *Proceedings of the 8th International Conference on Educational Data Mining* (pp. 250-257). Madrid, Spain.

Gillani, N., Yasseri, T., Eynon, R., & Hjorth, I. (2014). Structural Limitations of Learning in a Crowd: Communication Vulnerability and Information Diffusion in MOOCs. *Scientific Reports*, *4*.

Graesser, A. C., & McNamara, D. S. (2011). Computational Analyses of Multilevel Discourse Comprehension. *Topics in Cognitive Science*, *3*(2), 371-398.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing Multilevel Analyses of Text Characteristics. *Educational Researcher*, *40*(5), 223-234.

Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition*. Cambridge, U.K.: Cambridge University Press.

Knowles, J. (2013). Getting Started with Mixed Effect Models in R. Retrieved from: https://www.jaredknowles.com/journal/2013/11/25/getting-started-with-mixed-effect-models-in-r

Snow, C. E. (2002). *Reading for Understanding: Toward a Research and Development Program in Reading Comprehension*. Santa Monica, CA: Rand Corporation.