An Exploratory Study of Automated Clustering of Themes to Identify Conceptual Threads in Knowledge Building Discourse

Gaoxia Zhu, Leanne Ma, Andrew Toulis, and Monica Resendes
gaoxia.zhu@mail.utoronto.ca, leanne.ma@mail.utoronto.ca, andy.toulis@mail.utoronto.ca,
monicaresendes@gmail.com
University of Toronto

Abstract: In this study, we adopted Jaccard index and tf-idf without stop words to automatically cluster the ideas students discussed in on an online knowledge building platform called Knowledge Forum. We visualized the clusters, provided keywords that most represent the context of each cluster and compared the generated themes with manual coding themes. The results suggest that most of the generated themes were consistent with human coding results.

Introduction

Knowledge Building advocates that students take collective responsibility for continually improving ideas and pursuing more coherent explanations as a community (Scardamalia & Bereiter, 2014). During Knowledge Building, students generate diverse ideas, build onto each other's ideas, and introduce new ideas to their community both face to face and in Knowledge Forum-a software environment developed to support Knowledge Building practice (Scardamalia, 2004). Developing an understanding of the themes that a class works on is the starting point of understanding the frontier of community knowledge.

Methods to identify meaningful semantic themes have been a focal area of research in CSCL (e.g., Suthers, Lund, Rosé, Teplovs, & Law 2013). In this study, we suggest that text classification approaches, which extract and represent important information from documents, have the potential to help identify broad themes in online discussions (Mu, Stegmann, Mayfield, Rosé, & Fischer, 2012). The Jaccard index also referred to as Intersection over Union, has been widely used for comparing the similarity between samples in automatic classification, citation analysis, information retrieval and so forth (Hamers et al., 1989). To help capture conceptual threads in students' Knowledge Forum discussions, we developed a note clustering tool adopting the Jaccard index. This tool creates automated visualizations of sentences with overlapping keywords as clusters, and we created a conceptual label for each thematic cluster. In detail, we explored whether automatically generated themes were consistent with human coding results of conceptual threads in the student discourse.

Methods

The dataset analyzed in this study consists of 298 Knowledge Forum notes generated by grade 1 students (11 boys, 11 girls). Over the span of three months, students engaged in Knowledge Building discourse about the water cycle. Two researchers coded the Knowledge Forum notes into 15 conceptual threads (i.e., a group of notes which aim to address the same thematic issue). The manually coded conceptual threads were compared with the automated clusters generated by the note clustering tool used in this study.

The text classification processes mainly consist of five steps: 1) We manually spell-checked all the Knowledge Forum notes given the difficulty of automatically correcting the notes written by junior students. 2) We segmented the notes into sentences based on punctuations via the NLTK sentence tokenizer (Bird, Klein, & Loper, 2009). Then symbols, stop-words, and small sub-words were removed from the sentences (Patel, & Shah, 2013). Also, since most notes students wrote were related to water, we removed "water" from the analysis to achieve a clearer picture of other themes. 3) The lowest frequency for a word to be included into the analysis is tow since a keyword needs to appear at least twice to form a connection with other sentences. 4) We connected similar sentences together to form a network using Jaccard index. The metric we used for similarity threshold is representing the number of intersection keywords across sentences out of all words used in the union of keywords between any two sentences. We set 3/5 as the threshold. 5) We performed clustering on the network formed. We use the Louvain Method (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008) for clustering nodes into what are known as communities in the network sciences literature.

We visualized the networks formed under this procedure using visNetwork, a package in R. The notes are represented using red dots and were connected using red lines. In order to summarize the context of each cluster, we extracted the top words used by notes within each cluster using the tf-idf method which highlights top words used by the cluster and removes highly frequent words used across the whole dataset. Each cluster summary was marked using a blue dot, and the size of the blue dot indicates the popularity of the cluster – the more sentences (red lines) in a cluster, the bigger the dot (as shown in Figure 1).

Results

393 sentence units were segmented out of the dataset, and 136 keywords were kept in analysis. Figure 1 shows 16 clusters were formed by the tool. The two researchers identified 15 conceptual threads. Here, we qualitatively matched the two sources of clusters—the ones displayed outside of round brackets were generated by the tools while the ones inside the brackets were identified by the researchers: *water evaporates when it is hot (why does water evaporate, why can't you see water vapour); evaporation makes clouds and rain (how does water vapour go back into water, rain, why does the earth need clouds/water); clouds block vapor (where does water vapour go if there are no clouds); clouds are light so they float (how does water vapour float, clouds' weight, how can water be so light); the atmosphere stops clouds to go to space (the atmosphere); the color of clouds (clouds' colour); water freezes when the weather is cold (ice); and meteorites hit the earth (where did water come from, can you make water).* All the human coded themes were extracted by the note clustering tool except "groundwater." Possible reasons for why this theme was not picked up were that students did not use enough overlapping keywords when discussing this topic or the proportions of overlapping words they used in sentences did not meet the chosen threshold.

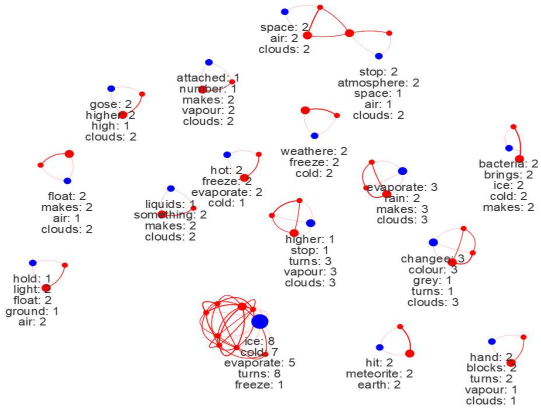

Figure 1. Automated clusters generated by the note clustering tool.

Discussions and conclusion

We see this method as an option to speed up the manual process of connecting notes together. We imagine people applying this network method to a raw dataset to jump-start the connection process for students, summary process for teachers and analysis process for researchers. We noticed that clusters may represent the same theme due to different keywords used, different ways students wrote their ideas, and different combinations of ideas. For instance, the two cluster context summaries "ice, cold, evaporate, turns, freeze" and "whether, freeze, cold" are both related to the theme of "ice." We also noted that the assumption of intersection of words misses out on words that are synonyms, such as "cold" and "freezing." Ideally, we would consider these words to be the same unit and hence count it in our intersection, and only count once in our union. We are working on a method currently for this as future work.

References

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc."

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10), P10008.

Hamers, L. et al. (1989). Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula. *Information Processing and Management*, 25(3), 315-18.

Mu, J., Stegmann, K., Mayfield, E., Rosé, C., & Fischer, F. (2012). The ACODEA framework: Developing segmentation and classification schemes for fully automatic analysis of online discussions. *International journal of computer-supported collaborative learning*, 7(2), 285-305.

Patel, B., & Shah, D. (2013, March). Significance of stop word elimination in meta search engine. In *Intelligent Systems and Signal Processing* (ISSP), 2013 International Conference on (pp. 52-55). IEEE.

Scardamalia, M. (2004). CSILE/Knowledge forum®. *Education and technology: An encyclopedia*, 183-192.

Scardamalia, M., & Bereiter, C. (2014). Knowledge building and knowledge creation: Theory, pedagogy, and technology. In K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (2nd ed.), pp. 397-417. New York: Cambridge University Press.

Suthers, D. D., Lund, K., Rosé, C. P., Teplovs, C., & Law, N. (2013). *Productive multivocality in the analysis of group interactions*. Springer US.