Office of Open Learning Publications

Office of Open Learning

2009

Determining curricular coverage of student contributions to an online discourse environment: Using latent semantic analysis to construct differential term clouds

C. Teplovs

Nobuko Fujita
University of Windsor, nfujita@uwindsor.ca

Follow this and additional works at: https://scholar.uwindsor.ca/open-learningpub

Determining Curricular Coverage of Student Contributions to an Online Discourse Environment Through the Use of Latent Semantic Analysis and Term Clouds

Christopher Teplovs and Nobuko Fujita, University of Toronto, 252 Bloor St. W., Toronto, Ontario, CANADA, christopher.teplovs@gmail.com, nobuko.fujita@gmail.com

Abstract: This paper presents a new approach to mapping student contributions to curriculum guidelines through the use of Latent Semantic Analysis and information visualization techniques. A new information visualization technique – differential term clouds – is introduced as a means to make clear changes in semantic fields over time.

Introduction

As part of their commitment to collective cognitive responsibility (Scardamalia, 2002), students engaged in knowledge building often pursue deep understanding of the topics being studied. A concern expressed by teachers, parents, and other observers is one pertaining to curricular coverage: what evidence is there that students are covering the mandated curriculum? To date, time-consuming manual analysis of student contributions has been the principal method by which answers to that question have been determined.

Methods

Twenty-two nine- and ten-year-olds worked on a unit on optics for ten weeks. They contributed notes to a Knowledge Forum database during that time. Overall, 152 notes were contributed in six shared working views. Each of the views ("Colours of Light & Rainbows", "Grey Fur & White Snow", "Reflection & Absorption", "Mirrors", "Where Light Goes & How", and "Lenses & Sight") corresponded to a topic area within the optics framework.

After completion of work by students and teachers, the Knowledge Forum database was augmented with additional notes. A new note was created for each of the curricular outcome statements contained in the Grade 4 Ontario Curriculum Guidelines (1998). Because the optics area covered in this grade spans two curricular areas, notes were organized into two views, each corresponding to one of the two curriculum units covered by the class's work on light and optics. Latent semantic analysis was use to create the semantic space and the cosines between all notes in the database were calculated. Only those exceeding 0.4 were retained because of the tenuousness of semantic relationships at cosine values below this threshold.

The Knowledge Space Visualizer (KSV) is a visualization tool capable of extracting notes from multiple views to aid the user in finding relationships amongst disparate notes (Teplovs, 2008). Participant contributions are represented as nodes that are linked via edges that correspond to explicit (e.g. "this note build on", "this note annotates", "this note references") or implicit (i.e. the content of these notes is similar) semantic links. Notes were extracted from each of the six shared working views in turn and juxtaposed with notes extracted from the two curriculum views. The KSV was used to arrange the notes into two lines as shown in Figure 1. The upper line of notes are the curriculum outcome statements. The lower line represent the student contributions. Edges indicating semantic similarity join the nodes. This arrangement of notes optimizes visualization of semantic connections between the two sets, while minimizing the distraction of linkages amongst notes within each set.
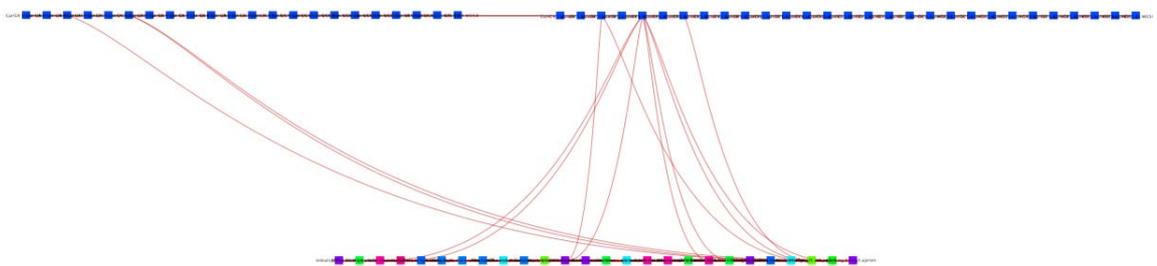
Figure 1: Visualization of curriculum outcome statements (upper horizontal line) and student contributions (lower horizontal line). There are 5 curriculum outcome statements linked to student notes.

Once arranged in this fashion the number of linked curricular items was recorded. To help determine where in the ten-week unit the majority of curricular overlap occurred the date range was adjusted to show only early, middle, and late contributions and the number of linked curricular items was recorded for each setting.

The process of visualizing and counting semantic linkages was repeated for cosine values of 0.9, 0.8, 0.7, 0.6, 0.5, and 0.4. High cosine values referred to highly specific linkages. Lower cosine threshold values represented looser linkages. Linkages below a cosine threshold of 0.4 were deemed too tenuous to merit a claim that the contents were semantically related. Finally, an independent (human) rater assessed whether notes showed evidence of mapping to curriculum guidelines.

To complement the semantic linkage analysis and to provide insight about the content of the participants' contributions a more detailed analysis that focused on the semantic fields produced by the participants was conducted. Term clouds, also known as tag clouds, are a relatively recent innovation. They consist of groups of words, often ordered alphabetically, whose fonts are scaled in size according to frequency of occurrence. The recent emergence of text clouds as a legitimate visualization technique holds promise for investigating the nature of online discourse by allowing the investigator to get a picture of what is happening in the discourse space. In other words, text clouds are useful visualizations for semantic fields. A shortcoming of traditional term clouds is that they do not do a particularly good job of showing semantic field growth. There are at least two reasons for this: there is no chronological aspect, and there is no sense of change vs. the previous term cloud (i.e. a sense of "delta"). The use of font sizing to scale the terms in the cloud is also questionable, as it merely highlights frequently used term rather than highlighting changes over time.

To address these shortcomings, a new graphical form of tag clouds was used. Small multiples (Tufte, 1990) were used to present a chronology of semantic field changes. In this technique, the term clouds reflect the changes between two other term clouds, and the frequencies are indicative of the changes in term frequency rather than the absolute numbers.

Results & Discussion

Overall, none of the student contributions mapped onto curriculum outcome statements at the highest similarity threshold (cosine $\geq$ 0.9). This is likely because of the very different nature of student discourse about a particular topic and the statement about the topic to be covered as gleaned from the government-produced curriculum documents. As the threshold for semantic similarity was lowered more curriculum outcome statements were considered sufficiently similar to the student contributions that a link indicating overlap was included, and that the specific curricular point was considered "covered". At the most inclusive threshold level (cosine $\geq$ 0.4) many, but not all, curriculum outcome statements were identified as having been covered.   The highest reliability was achieved at cosine values between 0.4 and 0.5.

More detailed analysis of what, exactly, participants were contributing to the discourse space is provided through the use of differential term clouds.   Figure 2 shows a sample differential term cloud.

absorbs (2) black (5) book* (2) colours (2) design* (2) evidence (1) experiment (2) flashlight (3) glass (1) green (3) heavier* (2) hits* (2) information (1) learned (1) lens (1) light (12) purple* (3) rainbow (2) red (3) reflect (1) shine (1) theories (1) theory (2) understand (1) water (3) white (1) yellow (1)

Figure 2: Sample differential term cloud. New terms are marked with an asterisk (*), curricular terms are underlined.

The term clouds can be examined to gain an understanding of what sorts of topics are being covered in the discourse space. The use of differential term clouds, rather than cumulative or sequential term clouds, highlights those terms that have been featured in a particular time period. Simply put, examination of the term cloud helps answer the question: "what's new this week". Of course, any size of time slice can be used; weeks were used as a convenient duration in this study. Large differential term clouds are indicative of significant changes in the focus or content of the discourse from which the term clouds are generated. The differential term clouds can be further summarized by looking at the breakdown on new, unique, and curricular terms as shown by the example in Figure 3.

A challenge for the use of term clouds in general is how to deal with misspellings and alternative word forms. In some cases it is clear that the misspelling should be corrected before the analysis. In others, such as the difference between "glass" and "glasses" the decision to combine them is less clear.

Conclusions

Latent semantic analysis (LSA) can be used to determine semantic similarity of documents. One application of LSA is to determine the semantic similarity of discourse from an online environment and descriptions of

curriculum to be covered. Although the nature of the discourse differs in terms of purpose and genre, meaningful comparisons can still be made. Term clouds generated by determining the differential contents of two term clouds that represent different time slices of the discourse can be used to provide additional information about the exact nature of the shifts in the discourse.
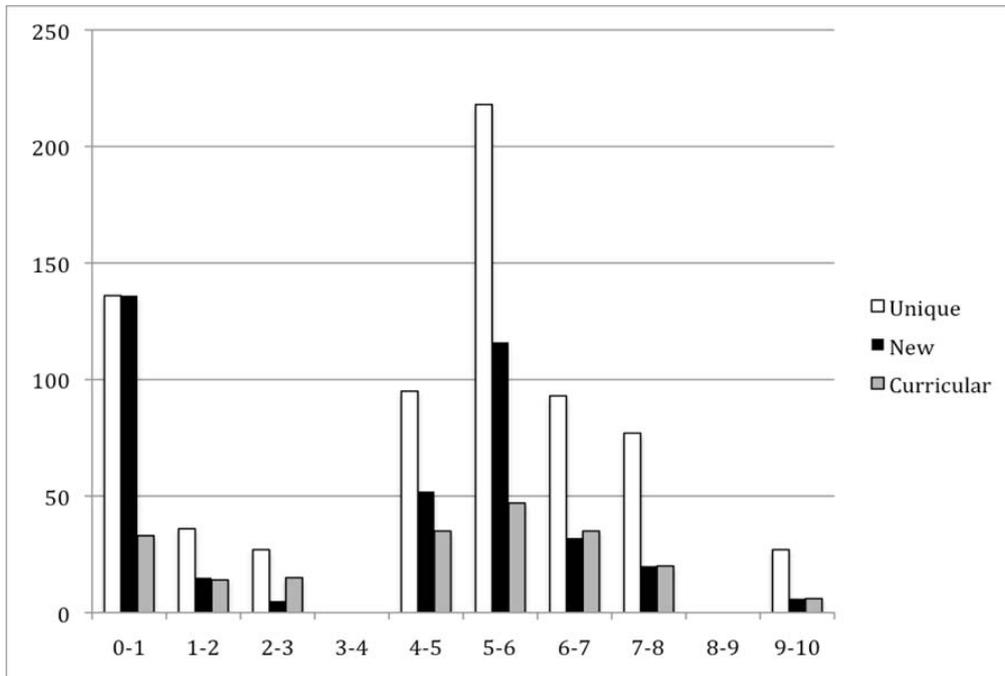

Figure 3: Week-to-week changes in unique, new, and curricular words in differential term clouds.

References

Ontario Ministry of Education and Training. (1998). The Ontario Curriculum Grades 1-8: Science and Technology. Toronto, ON: Ministry of Education and Training.

Scardamalia, M. (2002). Collective cognitive for the advancement of knowledge. In B. Smith (Ed.), *Liberal education in a knowledge society* (pp. 67-98). Chicago: Open Court.

Teplovs, C. (2008). The Knowledge Space Visualizer: A tool for visualizing online discourse. In G. Kanselaar, V. Jonker, P. A. Kirschner & F. Prins (Eds.), Proceedings of the International Conference of the Learning Sciences 2008: Cre8 a learning world. Utrecht, NL: International Society of the Learning Sciences.

Tufte, E. (2000). *Envisioning information.* Cheshire, CT: Graphics Press.