JOURNAL OF LEARNING ANALYTICS

(2017). Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *Journal of Learning Analytics*, *4*(3), 76–101. http://dx.doi.org/10.18608/jla.2017.43.5

Promising Ideas for Collective Advancement of Communal Knowledge Using Temporal Analytics and Cluster Analysis

Alwyn Vwen Yen Lee
Centre for Research and Development in Learning (CRADLE@NTU)
Nanyang Technological University, Singapore
alwynlee@ntu.edu.sg

Seng Chee Tan
Centre for Research and Development in Learning (CRADLE@NTU)
Nanyang Technological University, Singapore

ABSTRACT: Understanding ideas in a discourse is challenging, especially in textual discourse analysis. We propose using temporal analytics with unsupervised machine learning techniques to investigate promising ideas for the collective advancement of communal knowledge in an online knowledge building discourse. A discourse unit network was constructed and temporal analysis was carried out to identify promising ideas, which are improvable perceptions of significant relevance that aid in the understanding of discourse context and content. With the aid of a degree centrality–betweenness centrality (DC-BC) graph, more promising ideas were discovered. An additional analysis using multiple DC-BC graph snapshots at different discourse junctures illustrates the transition of these promising ideas over time. Machine learning in the form of *k*-means clustering further categorized promising ideas. Cluster centroids were calculated and represented the foci of discussions, while the movement of discourse units about cluster centroids reflected how ideas affected learning behaviours among the participants. Discourse units containing promising ideas were qualitatively verified. Overall, the results showed that the implementation of temporal analytics and clustering provided insights and feedback to users about idea-related processes in the discourse. The findings have implications for teachers, students, and researchers.

Keywords: Temporal analytics, machine learning, cluster analysis, promising ideas, idea analysis, knowledge building discourse

NOTES FOR PRACTICE

- Analysis of ideas and related processes is critical for collective knowledge advancement in knowledge building discourse but is challenging due to large and complex discourse data.
- This paper contributes to the field of learning analytics by proposing a method that combines temporal analytics and unsupervised machine learning to analyze promising ideas and investigate idea mobility in discourse.
- Findings show that temporal analysis bridges the gap between individual analyses of discrete events to provide a broader picture of online discourse that is complementary to machine learning techniques and provides insights into idea-centric discourse.

(2017). Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *Journal of Learning Analytics*, *4*(3), 76–101. http://dx.doi.org/10.18608/jla.2017.43.5

1 INTRODUCTION

In an educational setting, individuals interact and share ideas, collaborate and build their understanding of the world, and through this process, advance communal knowledge through discourse. With increasingly accessible discourse forums and tools, educational institutions, universities, and schools are encouraging the student population to engage in online discourse. These discourse data are often archived but mostly left untouched. Discourse analysis could help to provide an understanding of language beyond literal usage, and further detailed analysis can even inform the design and productivity improvement of knowledge creation (Chiu & Fujita, 2014a). Content analysis of asynchronous discussions can help to detect cognitive presence in online discussions (Kovanović et al., 2016), and institutions can also learn new information about student learning to enhance educational conditions related to student success (Baer & Campbell, 2012). The trends and insights obtained from such analyses could be used for improving teaching and learning efficacy. Therefore, a growing number of researchers are investigating student discourse to understand classroom interactivity and the level of participant understanding.

One challenge, however, is the immeasurably large amount of content and data within an online discourse. Increasingly complex datasets are constantly being created on discourse platforms, but traditional analytical methods such as word counts and frequency lists cannot harness the full potential of these data. In fact, with multiple sources of data and features to choose from, researchers are focusing on specific types of context, data, and instruments to determine various impacts towards teaching and learning. When dealing with discourse data, learning analytics are often used, such as statistical discourse analysis (SDA; Chiu & Fujita, 2014b) and discourse-centric learning analytics (DCLA; Knight & Littleton, 2015). Researchers have also used semantics in identifying topic specificity in online discussion forums, through probabilistic topic modelling with semantic visual analytics (Sun, Zhang, Jin, & Lyu, 2014; Hsiao & Awasthi, 2015). Software-based tools, such as the Idea Thread Mapper tool (Zhang, Chen, Tao, Naqvi, & Peebles, 2014), were created to support collaborative reflection for sustained knowledge building. A Promising Ideas tool (Chen, Scardamalia, & Bereiter, 2015) was also developed to aid in the advancement of knowledge building discourse through judgements of promising ideas.

For analysis of discourse, temporal considerations are important in gaining deeper understanding of the processes of learning over time, and yet they continue to be understudied in educational research (Piety, Hickey, & Bishop, 2014; Chen, Wise, Knight, & Cheng, 2016). Nonetheless, there is a growing interest in temporal analytics among researchers and developers. Recent international conferences such as the International Conference on Learning Analytics and Knowledge (LAK) and workshops (Knight, Wise, Chen, & Cheng, 2015; Chen et al., 2016) have increasing focus on the creation of temporal analytics tools that could help users make sense of educational temporal data. This could be attributed to the emergence of temporal analytics and machine learning techniques able to provide alternative approaches for discourse analysis. These alternative methodologies are able to process multi-dimensional data, analyze data against a continuum such as time, and provide deeper insights into what is going on in a discourse. Examples of the methodologies include machine learning algorithms (Garrard, Rentoumi, Gesierich, Miller, & Gorno-Tempini, 2014), clustering techniques with part-of-speech (POS) tagging, natural language

(2017). Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *Journal of Learning Analytics*, *4*(3), 76–101. http://dx.doi.org/10.18608/jla.2017.43.5

processing (NLP) capabilities (Owoputi et al., 2013), and a range of temporal analytics (Molenaar, 2014). By considering the knowledge that grows in shared spaces (such as online forums) as community knowledge (Scardamalia & Bereiter, 2006), some of the current methods and tools (e.g., SNAPP; Bakharia & Dawson, 2011) are able to represent this knowledge and relationships through near real-time analysis of discussion forums. In addition to temporal analytics, unsupervised machine learning could help to gain deeper understanding and establish baseline behavioural profiles to find meaningful anomalies. Researchers can use a combination of these methods to improve the effectiveness of understanding through visualizations, and make it easier for students to monitor and advance communal knowledge.

This study calls attention to the use of temporal analytics and machine learning in online knowledge building discourse to analyze promising ideas from various participants in the discourse, and to investigate the types and movement of these ideas. This study contributes to the field of learning analytics by offering a new method that combines the use of both temporal analytics and unsupervised machine learning, specifically cluster analysis. As the research of idea analysis in discourse is rather new to learning analytics, we believe there is still work that can be done to help teachers and students understand further and continue their efforts in encouraging creative work for sustained idea development in discourse. Recent developments from Chen, Scardamalia, and Bereiter (2015) and Lee and Tan (2017) initiate baseline research on promising ideas and temporal discourse analysis. We seek to improve on the current work to provide a clearer analysis and visualization of ideas in discourse. More so, the method proposed in this paper can help determine the types and movement of ideas initiated by teachers and students during and after discourse, which will likely assist participants to monitor their own effort in sustaining creative work and improvement of ideas in knowledge building discourse. The overarching research question guiding this study is this: "How can the application of temporal analytics and machine learning techniques be used to identify and understand the movement of ideas that are promising to the collective advancement of communal knowledge in an online knowledge building discourse?"

2 LITERATURE REVIEW

2.1 Need for Analysis of Promising Ideas in Knowledge Building Discourse

Locke (1836) initially used the word *idea* to represent the most basic unit of thought, and refer to ideas as immediate objects of perception that are interesting to a person since the ideas point beyond themselves. Modern definitions consider ideas as "transcendent entities that are a real pattern of which existing things are imperfect representations" (Merriam-Webster, n.d.). Ideas can also be treated as real things, as objects of inquiry and improvement in their own right (Scardamalia & Bereiter, 2003). Hence, an idea is not just a unit of thought, but more of what it can achieve, such as the provision of epistemic function to represent something else and the ability to improve beyond itself. In discourse, ideas often mean something pictured in mind, such as an emerging development of a concept, an evolving process, or a way of explaining phenomenon. These ideas are often represented in inquiries, statements, or claims in spoken discourse, and are crafted as part of forum posts in online discourse.

(2017). Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *Journal of Learning Analytics*, *4*(3), 76–101. http://dx.doi.org/10.18608/jla.2017.43.5

In a traditional instruction-based classroom that treats learning as the acquisition of knowledge (Sfard, 1998), the factual nature and pre-assigned authoritative sources of content tend to lead students to accept presented ideas as facts, leaving little room and time for sharing and improving ideas through classroom discussion. There are, in fact, other approaches to learning, such as learning as participation (Sfard, 1998) or learning through knowledge creation (Paavola & Hakkarainen, 2005). By allowing the student community to engage in discourse, students have the chance to share and improve ideas, and create new knowledge through in-depth discussion.

We chose knowledge building (Scardamalia & Bereiter, 2003) as an approach to knowledge creation in education. We implement knowledge building as a pedagogical approach because it allows us to leverage the natural learner capability of idea generation for collaborative improvement of ideas. With this, teachers can maintain student engagement in creative work to support processes of idea improvement. Discourse is an important medium that plays a creative role in encouraging the improvement of ideas (Lakatos, 1970), and the productive use of the principle of "improvable ideas" through inquiry (Scardamalia, 2002), argument, and debate can lead students to treat every contribution and idea in discourse as potentially improvable, and eventually develop better ideas collaboratively. In such a learning environment, the principle of idea improvement (Scardamalia, 2002) is crucial for students to acknowledge knowledge gaps, to navigate among emergent themes of inquiry from multiple sources of inputs, and to work collaboratively (Zhang, Scardamalia, Reeve, & Messina, 2009). When students engage in knowledge building discourse to share information and seek solutions to their own problems, they are also improving on one another's ideas through the discourse; such a process culminates in elevating the community's level of understanding. Finding a way to detect ideas and analyze idea-related processes is, therefore, critical to any knowledge building discourse.

At the initial stage, most ideas in a discourse are represented in preliminary forms with uncertain prospects (Chen, Scardamalia, & Bereiter, 2015), and the majority of ideas would be eliminated as the learners increasingly focus on those worth pursuing. Therefore, to help achieve a higher level of communal understanding, we could watch for ideas in a discourse that are able to take on additional meaning in the context of creative or design thinking (Martin, 2009); in other words, ideas that are improvable and capable of moving the community forward. These are ideas with *promisingness* (Chen, 2014), which are important ingredients of knowledge creation (Gardner, 1994). Evaluation of promisingness has been previously conducted in studies of other areas of expertise, such as the evaluation of expert writers based on promisingness pattern recognition (Bereiter & Scardamalia, 1993). More recent work has focused on intentionally leaving the judgement of promisingness to knowledge builders (Chen, 2017), as setting a fixed definition for promisingness using computational methods is admittedly a challenging prospect even with current approaches and technology.

Where the focus is on the advancement of communal knowledge within an online knowledge building discourse, promising ideas are considered and evaluated equally from perspectives of all participants, be it students or teachers, depending on the learning objectives and instructional goals respectively. The usage and make-up of participant vocabulary in discourse are often construed as the ability to which

(2017). Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *Journal of Learning Analytics*, *4*(3), 76–101. http://dx.doi.org/10.18608/jla.2017.43.5

participants can express their ideas, so the network of keywords forming a semblance of ideas and opinions is largely representative of the extent to which participants are able to enunciate the relevance and potential of their ideas to others. In this regard, we consider that a fairly accurate assessment of idea promisingness, reflective of the content and intent in discourse, can be conducted from an analysis of the vocabulary make-up of the discourse, along with the identification and measurement of connections between groupings of similar ideas found in the discourse. These analyses can be determined through the application of learning analytics and techniques, like temporal analytics and machine learning, to capture and reflect the perceptions and judgements of idea promisingness in discourse, and further identify ideas promising to the collective advancement of communal knowledge.

2.2 Using Temporal Analytics and Machine Learning in Idea Analysis

In this paper, temporal analytics and machine learning techniques, specifically cluster analysis, are used for idea analysis.

2.2.1 Temporal analytics: Temporal network analysis using keywords from discourse

To gain deeper insights into the process of online discussions, it is necessary to bridge the gap between individual analyses of discrete events in order to provide more information to illuminate the context of the event. This requires an analytical process that can connect numerous minute details at the micro-level with the underlying theory that operates on the macro-level (Mercer, 2008). For example, clickstream data contains copious amounts of information that is, by itself, meaningless. At the micro-level, the data can be used for determining the relationship between goal achievement in MOOCs and behavioural clickstream analysis (Wilkowski, Deutsch, & Russell, 2014) but is unable to predict course completion rate. However, by recognizing similar data as temporal patterns, Chen, Haklev, Harrison, Najafi, and Rolheiser (2015) were able to provide additional insights into student actions, thus showing the potential for determining predictive actions and potential interventions using temporal analytics. Further, the process of teaching and learning has a long-term trajectory that cannot be understood only as a series of discrete educational events (Mercer, 2008). The “code-and-count” method that aggregates discrete events over time tends to ignore fine-grained details (Suthers, 2006). Hence, temporal analytics has been suggested as the way forward for analyzing learning processes (Reimann, 2009; Schoor & Bannert, 2012), but this form of analysis is still undervalued and not studied in detail (Mercer, 2008). Through temporal analysis, characteristics of data pertaining to the discourse participants are illustrated, and this information can be used for navigating through series of events that unfold over time. In this study, we leverage temporal analytics to gain a deeper analysis of promising and improvable ideas contributed by discourse participants.

The understanding of ideas in a discourse is, nevertheless, a challenging process, especially so when the analyzed content consists of unstructured textual data. The text analysis process is complex and heavily dependent on context to ensure accuracy of understanding. Effort has been made to find meaningful patterns in text through key phrase matching or visualizations of category patterns (Rosé et al., 2008), while machine learning approaches towards NLP have also been used recently for semantic analysis and

(2017). Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *Journal of Learning Analytics*, *4*(3), 76–101. http://dx.doi.org/10.18608/jla.2017.43.5

automatic text analysis in social media (e.g., Gruzd, Paulin, & Haythornthwaite, 2016). A way to circumvent this problem is the use of keywords. Keywords in a discourse indicate the conceptual understanding of what is being discussed and they can be used as proxy indicators of the ideas being generated and discussed by the discourse community. Commonly used keyword-based technologies can provide researchers with lexical foci of any structured text, such as revealing the hegemonic nature of discourse using a frequency list (Baker, 2006). Keyword-based technologies have been used for other purposes, such as analyzing student engagement (García & Caplan, 2014), determining trends in social networks (Nayyar, Hashmi, Rafique, & Mahmood, 2016), and conducting social network analysis of discourse (Oshima, Oshima, & Matsuzawa, 2012). Keyword-based technologies can thus also be harnessed for temporal analysis to approximate ideas and context in discourse. An example used in previous studies (Oshima et al., 2012; Lee, Tan, & Chee, 2016) is the analysis of graphs based on bipartite relationships that reveal the associations between keywords, discourse units, and participants within a discourse. These developments show that the usage of groups of keywords by various participants can represent some semblance of ideas shared across different discourse units.

2.2.2 *Machine learning: Cluster analysis using k-means clustering*

Machine learning is a subfield of computer science that develops the ability of computers to learn without being explicitly programmed (Samuel, 1959). The advantage of machine learning techniques and algorithms over traditional methods of approaching large datasets is the iterative learning process that uses previous computations to produce more reliable and robust results, which are then reused in subsequent iterations for repeatable and better decisions. Even though machine learning is not a new field in computer science, the application of machine learning is, however, a recent and significant foray into the field of educational technology, garnering significant interest from researchers in automating analysis to increase the efficacy of methodology and improve pedagogy.

We use cluster analysis (clustering), which originated in the field of psychology (Bailey, 1994) as one of the approaches for machine learning. Many clustering algorithms are being used for different purposes and datasets (Estivill-Castro, 2002). The common approach towards clustering is the unsupervised form of machine learning that does not require any training sets, as compared to the supervised approach that trains clustering algorithms to achieve desirable clusters (e.g., Finley & Joachims, 2005). The unsupervised approach partitions sets of unknown inputs into groups, which can be viewed as a process for grouping and labelling sets of raw data that have no inherent belonging to any group. It is, therefore, useful for providing insights into unclassified data that are otherwise difficult to understand.

Other than the benefits of using clustering as an unsupervised form of machine learning, clustering has been used to help researchers develop profiles grounded in learner activity (Antonenko, Toy, & Niederhauser, 2012), such as initiation and sequence of events. Clustering plays a significant role in discourse analysis, as analyzed content of similar discourse features can be recognized and clustered in order to identify profiles and types of ideas. Among different clustering algorithms and models, a cluster hierarchy can be generated top-down using a technique called divisive clustering. Clusters are split using a flat clustering algorithm to create a flat set of clusters without explicit structures that relate clusters to

(2017). Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *Journal of Learning Analytics*, *4*(3), 76–101. http://dx.doi.org/10.18608/jla.2017.43.5

each other; this algorithm is more efficient when a fixed number of top levels is chosen instead of constructing the complete hierarchy down to the individual leaves. We chose the efficient flat clustering algorithm *k*-means to analyze characteristics of ideas during discourse analysis, mainly because of its ability to reduce massive data to simple centroids and split up large spaces of data into smaller disjointed sub-spaces called *Voronoi* cells. The *k*-means clustering algorithm can present massive amounts of discourse features and the centroids of clusters for end-users to focus on, providing easy to recognize and understand groupings. More importantly, implementation of the method only requires the number of expected clusters "k," which can be estimated before processing and can be adjusted for future iterations if the initial results are not meaningful for the study.

This study aims to contribute to learning analytics research in knowledge building discourse by 1) introducing the application of analytical approaches that can assist discourse participants in understanding the types and movements of ideas, and 2) providing methods for discourse analysis to enhance visibility of promising ideas.

3 DATA AND METHODS

The content and usage of language within discourse change over time, and these developments can be observed by focusing on the temporal dimension of the discourse. Other than tracking the frequency of keyword usage, we can also focus on the temporal trends of network measures. Prior work such as the usage of the I^2A methodology (idea identification and analysis; Lee et al., 2016) in knowledge building discourse has shown that promising ideas can be identified and categorized to provide an approximate understanding of idea types within the communal discourse. In essence, the I^2A methodology uses social network analysis and measurement coefficients of chosen keywords to identify and trace the evolution of ideas within a discourse space. In the previous study (Lee et al., 2016), the I^2A methodology was used to identify promising ideas to teachers and instructors using keywords chosen by teachers to analyze discourse. Another study (Lee & Tan, 2017) investigated ideas promising to students by using keywords extracted from online student discourse. Results from both studies were validated using qualitative analysis and the participants in both were able to determine the promisingness of ideas based on their own perceptions and opinions. For this paper, we deployed an improved methodology built on I^2A, consisting of temporal analytics and clustering, to provide a more detailed analysis and clearer visualization of both student and teacher inputs in discourse. By understanding and using network measures generated from the inputs in discourse, temporal analytics were implemented to explain the contrast between the ideas originating from student and teacher inputs, and visualizations were developed to show the transition of ideas and how they were affected over the period of discourse.

3.1 Dataset and Settings

This paper uses a dataset from the discourse of a graduate-level class in an educational institution, which consisted of 13 in-service teacher participants and two instructors who engaged in the discourse over 13 weeks. The key focus of their discussion included the basic principles of knowledge building and how the

(2017). Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *Journal of Learning Analytics*, *4*(3), 76–101. http://dx.doi.org/10.18608/jla.2017.43.5

knowledge creation model can be applied in future for their own learning or in classes that they would instruct. The discourse was conducted on the Knowledge Forum (Scardamalia, 2004), a knowledge building environment that supports online discourse. A total of 281 Knowledge Forum notes were generated by the community and archived during the period of discourse. In the next section, we report on the procedures and measures that use this anonymized data.

3.2 Methods and Measures

3.2.1 Identifying keywords from discourse to form a semblance of ideas

To determine promising ideas from discourse, data was retrieved from the knowledge building discourse archived on Knowledge Forum. To facilitate the process of managing large amount of discourse data, the discourse was broken down into individual conversation turns, and these turns, also known as "discourse units" or DU, can be inspected using visualization support from discourse analysis tools (Oshima et al., 2012). A DU for this study also refers to a Knowledge Forum note that contains content and ideas contributed by the note's author. Each note, consisting of content that the author contributed to the discourse, was anonymized with a code.

Keywords, detected using text-mining techniques, are the basic units of analysis in this study; their presence in discourse units indicates a partial resemblance of ideas. In the previous study (Lee et al., 2016), instructor-provided keywords were used because they represented the intended learning objectives of the instructor. However, instructor-provided keywords do not necessarily reflect participant views and opinions or their interest within a discourse. We used a method of extracting keywords from discourse to investigate the participant foci of discussion. We used the text-miner SOBEK, based on the work of Schenker (2003) and adapted for educational purposes as a text mining tool (Reategui, Epstein, Lorenzatti, & Klemann, 2011). SOBEK mines textual data and generates related conceptual keywords by identifying relevant, recurring terms to be presented in a graphical manner. SOBEK does not use training data and the results are not influenced by external inputs. Text-mined keywords are representative of ideas that transpired in the knowledge building discourse; results from SOBEK are a source of keywords from the textual discourse and context. For greater accuracy, SOBEK uses a built-in thesaurus that filters out common words, such as noun markers (e.g., determiners such as a, an, the, this), pronouns (e.g., his, him, her), and words with similar meanings (e.g., student, students, pupil). Keywords identified using the text-miner are presented as a graph of nodes, with synonyms grouped together under a single node (keyword), size of nodes representing the frequency of usage, and connections between nodes representing relationships between keywords. The resulting graph then provides a semblance of ideas that can be found in the discourse. The results (see Figure 2) in section 4.1 show the list of conceptually relevant keywords that SOBEK mined from the study's textual discourse and the respective generated keyword graph.

3.2.2 Using network measures and temporal analytics to approximate promising ideas in discourse

Once the list of keywords was generated, the Knowledge Building Discourse Explorer (KBDeX; Oshima et al., 2012), a discourse analyzer, was used to generate networks based on bipartite relationships that

(2017). Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *Journal of Learning Analytics*, *4*(3), 76–101. http://dx.doi.org/10.18608/jla.2017.43.5

associate keywords, discourse participants, and discourse units on a single analytical platform. This paper places more focus on the relationship among the discourse units and keywords (rather than among participants) because it aims to identify promising ideas within the discourse units. The discourse unit network is, therefore, generated based on the relationships between keywords and discourse units, with the network subsequently being used for the calculation of conventional network measures, such as betweenness centrality (BC) and degree centrality (DC). We first explain the role of BC, used in the previous and current study, for understanding temporal trends and identifying promising ideas in discourse units, followed by the integration of DC with BC to provide deeper insights into the discovery of undetected promising ideas.

The BC measure is of interest as it indicates the degree of importance and connectivity that a discourse unit can provide to help connect ideas. Betweenness centrality, when measured in the context of discourse unit networks, is also an indicator of how well ideas can mediate important connections to ideas in other discourse units. A high BC value means that the ideas in the discourse unit are connected to many ideas in other parts of discourse hence are possibly promising and worth investigating.

The following example (Figure 1) illustrates our temporal analysis, using the dataset from Lee and Tan (2017). We used a two-phase procedure for the temporal analysis of the discourse units. The first phase involves the introduction of the idea and the peaking of BC value. The BC peak value reflects the highest point of interest by the community in the content of the analyzed discourse unit. Through qualitative analysis, the ideas within the DU can be examined to identify characteristics that lead to further discussion among the community (e.g., thought-provoking, novel, or disruptive ideas). The second phase of the analysis examines the impact and influence of the discourse unit's content on the community. Discourse units with ideas and content that do not sustain the community's interest over the long run tend to have a drastic drop in BC values over time. If the instructor or part of the community feel that the ideas and content of the discourse unit are potentially promising enough to advance communal knowledge, then extra effort is required to help support work in sustaining community engagement and interest. In this example, interest and sharing of ideas and content in DU8 waned slightly after the early peak in discourse, and there was a slight but steady decline in the BC trend until the end of discourse. Qualitative analysis revealed that DU8's content was sufficiently promising to sustain the community's interest throughout the rest of the discourse.

Although BC usage in prior studies was sufficient in showing the mediating role of a discourse unit in connecting ideas throughout the discourse, we included degree centrality as an additional measure for assessing network structure density and robustness. Degree centrality can be used to reflect the number of discourse units connected to a specific discourse unit. An increased degree centrality over time means that discourse units are connected to newer related information in other discourse units through interactions. It could also mean the ideas within the discourse unit are being reinforced by other students (as shown in the content of their discourse units). The discourse unit network thus becomes denser and more robust with more interactions in the network.

(2017). Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *Journal of Learning Analytics*, *4*(3), 76–101. http://dx.doi.org/10.18608/jla.2017.43.5

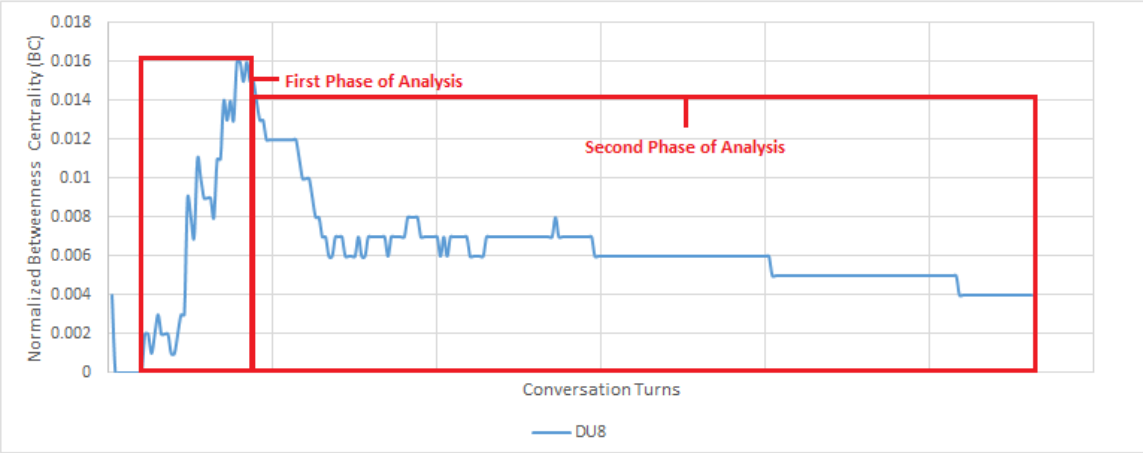

Figure 1. Analyzing the BC trend of discourse unit DU8 using a two-phase procedure.

A similar combination of the BC and DC measures was also used by Oshima, Oshima, and Fujita (2016) to distinguish epistemic actions for awareness of lack of knowledge in students. However, our goal of combining the BC and DC measures is to aid in the identification and visualization of promising ideas of instructor and student inputs in the discourse. When participants use keywords in sharing and exchanging information, they are trying to learn by creating meaningful links between normal communicative speech and usage of important keywords. Therefore, key ideas are central to discussions and for mediating thoughts and opinions; they can help generate newer ideas or improve current ones. The instructors do likewise on a smaller scale in helping to co-create knowledge, inject novel ideas, and provide guidance to students who require assistance, while maintaining class discipline in the online discussion environment. In short, examining BC and DC measures can provide additional information on the degree of sharing and level of communication by instructors and students within the discourse network.

3.2.3 Using variable space plot to understand idea mobility in discourse over time

Further value can be added to idea analysis by organizing the discourse data into a variable space plot to reveal relationships between discourse units. We first construct an overall view of all discourse units by plotting the BC and DC measures of individual discourse units onto a two-dimensional variable space plot. By representing the BC and DC values as x- and y-coordinates on a two-dimensional variable space, we can visualize and interpret the degree of idea promisingness in discourse units. We refer to the two-dimensional variable space as the DC-BC graph; the values of DC and BC are normalized for the discourse. The DC-BC graph has the added benefit of allowing all discourse units to be visualized side-by-side on the same plot, so that groups of discourse units in close proximity can be identified. Proximity indicates ideas of similar promisingness, since they possess similar DC and BC values. Discourse units with a high degree of promisingness possess relatively high DC and BC values; they are thus notably located in the top-right quadrant of the graph (such as in Figure 6). When newer discourse units are introduced into the discourse network, the promisingness of newly introduced ideas may be predicted by comparing the relative position of the new discourse unit with existing discourse units in the DC-BC graph.

(2017). Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *Journal of Learning Analytics*, *4*(3), 76–101. http://dx.doi.org/10.18608/jla.2017.43.5

Apart from the static visualization of discourse units on a DC-BC graph, multiple snapshots of the DC-BC plots over a period of discourse can be carried out to visualize the movement of ideas across time. These snapshots can be captured at any temporal juncture in a discourse, and the window of analysis can be situated between any two chosen temporal junctures, depending on the learning goal that the analyst or instructor wants to achieve. By monitoring a data point (representing measures of a discourse unit) on the DC-BC graph over time, the transition of ideas can be observed. The content and ideas in the discourse unit may have gained or lost relevancy throughout discourse, with changing degrees of promisingness to the community. We call the ease or difficulty of idea movement on the DC-BC graph "idea mobility." Discourse units with consistently high DC and BC values that do not significantly drift on the DC-BC graph over time often contain ideas shared and discussed, and therefore considered consistently promising to the community. By visualizing discourse units on the DC-BC graph and using temporal analytics, we are able to provide estimates of idea promisingness and reveal relationships of discourse units in close proximity over the period of discourse.

3.2.4 Conducting k-means clustering of discourse data to identify types of ideas in discourse

Even though understanding the movement of discourse units is important for presenting trends of ideas and movements of discussions, we can only estimate the promisingness of ideas and cannot identify the types of ideas based on movements and relative shifts. Therefore, we propose using clustering as an additional layer of analysis of the possible types of ideas that emerge from the discourse.

We use *k*-means clustering as a form of optimizing discourse unit clusters. When the number of clusters is fixed to a number "k" and data objects are assigned to the nearest cluster centre, clustering calculates minimum within-cluster squared distances between cluster members, and maximal inter-cluster distances from surrounding clusters. In essence, when we apply clustering to the DC-BC graph to form "k" number of clusters of discourse units, we are discovering how discourse units containing similar ideas can be grouped based on their DC and BC values. By assigning discourse units to different groups, it is easier to visualize how individual discourse units can be categorized according to the different types of ideas: promising, potential, and trivial (Lee et al., 2016). We differentiate the types of ideas using the following three factors: 1) relevancy to the community, 2) sustainable level of interest to the community, and 3) likely impact of the idea on discourse. There are likely three types of ideas found in the discourse. First, *promising ideas* are of great relevancy to the community and are able to sustain community interest; hence, they are worth pursuing and are likely to affect communal discourse. Second, *potential ideas* show some communal relevance, but interest in them is difficult to invoke or sustain, thus requiring scaffolding and intervention in order for the ideas to have some impact or influence on communal discourse. Third, *trivial ideas* contain little or no relevancy to the community and thus do not spark interest or affect discourse. In addition to the identification of idea types, the application of *k*-means also calculates the centroids of clusters. These centroids, also known as central vectors, are important indicators that visualize group-centric positions of idea groups within the whole discourse. Observations of discourse units and centroid movements throughout a discourse can reveal how the centre of discussion has shifted over time due to changes in learning behaviours among the participants.

(2017). Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *Journal of Learning Analytics*, *4*(3), 76–101. http://dx.doi.org/10.18608/jla.2017.43.5

To implement *k*-means for grouping ideas on the DC-BC plot, the number of clusters "k" has to be specified in advance, which is one of the limitations of *k*-means. The value of "k" must be at least 1, or not more than the value of maximum number of samples –1. Visual inspection of data points is one way to determine the number of clusters; prior studies and experience also play a part in helping to choose an appropriate "k" value. From previous studies (Lee et al., 2016; Lee & Tan, 2017), the categories of ideas in discourse were qualitatively verified after being identified using the proposed methodology. Findings have shown that the various patterns of BC trends tend to belong to certain groups of ideas. Therefore, when we consider clustering the same dataset on a variable space plot, we expect similar patterns to surface that can lead to the identification of three clusters representative of the three categories of ideas. We decided to start the study using k=3, since we expected to identify three categories of ideas (promising, potential, and trivial). Remember that the value of "k" can still be adjusted accordingly when the need arises. Of course, there will be some inherent subjectivity in the labelling process as we are conducting unsupervised learning with no training data and searching for an approximate solution. Even though three clusters are expected, labelling results of some data points may vary (e.g., run 1: promising idea, run 2: potential idea) over multiple runs of the clustering process. Therefore, this study only reports the best results of multiple iterations to reduce the local optimum problem and to show the final optimal results. A DC-BC graph that displays discourse units in three clusters, with calculated centroids for each cluster, can show the findings.

4 RESULTS AND DISCUSSION

To recapitulate, the following procedures were carried out:

1. Keywords were identified using the text miner SOBEK.

2. Keywords and notes were used to form a discourse unit network; discourse units containing promising ideas were identified using temporal analysis and visualization on the DC-BC graph.

3. Idea mobility in discourse was visualized on the DC-BC graph using multiple temporal junctures to represent transition of ideas over time. The window of analysis was chosen between two temporal junctures, namely the mid-discourse and end of discourse, because these milestones allowed us to conduct a meaningful investigation on promising ideas that emerged after some communal discourse has occurred and until the discourse has ended.

4. k-means clustering was applied to show the groupings of discourse units, categories of ideas, and cluster centroids, with temporal analysis showing the movement of discourse units, representing idea mobility of instructors and participants. The results are reported in the subsections that follow, along with related discussions on how temporal analytics and machine learning were used to uncover insights or explain a certain phenomenon during idea analysis.

(2017). Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *Journal of Learning Analytics*, *4*(3), 76–101. http://dx.doi.org/10.18608/jla.2017.43.5

4.1 Keyword Text Mined from Discourse

The knowledge building discourse was text-mined, and the resulting keywords were used as inputs for subsequent parts of the idea analysis. The identified keywords are indicators that represent ideas, which are in turn representative of main themes in the discourse. The identified keywords were "knowledge," "kb," "learning," "students," "knowledge building," "understanding," "discourse," "community," "idea," "based," and "information," in order of decreasing frequency. The keywords, displayed in a list (Figure 2), are linked based on Schenker's (2003) graph-theoretic technique.

We observed that participant usage of these keywords represents their inquiry of the concepts behind keywords and reflects the level of knowledge possessed by the participants. Considering that keywords are unique and unlikely to be repeated offhandedly by participants without careful thought or sufficient evidence of understanding, their wrong usage could be indicative of misconceptions or topics that the participants are uncertain of, and are still seeking clarification. The misinterpretations in discourse are corrected by sharing information, by being open to critique and listening to others, or through corrective guidance from the instructor. Instructors are more likely to use the keywords for prompting students, by guiding them towards understanding of a more complex process or concept.

Keywords from Discourse	
knowledge	kb
learning	students
knowledge building	understanding
discourse	community
idea	based
information	

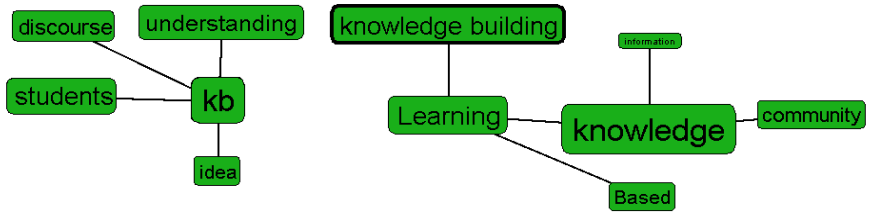

Figure 2. List of text-mined keywords and keyword graph from the discourse.

4.2 Discovering Promising Ideas and Idea Mobility

Using the keywords obtained through text mining, a discourse unit network based on bipartite relationships between discourse units and keywords was constructed using KBDeX. The BC values of individual nodes (notes by participants and instructors) in the discourse unit network were calculated over time, thus allowing us to form BC trends to understand the degree of mediation offered by discourse units at different conversation turns, or at different "time frames" of the knowledge building discourse. The resulting BC trends reflect the level of interest in ideas and content of discourse units, the engagement by the community in mediating knowledge, and the sharing of ideas among the community.

Instead of displaying hundreds of BC trends on a single graph, a few discourse units were selected and explained. Figure 3 shows four BC trends of discourse units containing content of high relevancy and ideas that attract communal interest, with three of the discourse units (DU8, DU18, and DU66) belonging to the participants and one discourse unit (DU54) belonging to the instructor. These discourse units exhibited higher than usual BC peaks, often after the initiation of the discourse unit. The content of discourse units

(2017). Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *Journal of Learning Analytics*, *4*(3), 76–101. http://dx.doi.org/10.18608/jla.2017.43.5

DU8 and DU18 were widespread and dominated the early stages of discourse, partially due to the small number of participants and limited variety of opinions. As the size of the discourse unit network grew over time, the respective BC trends exhibited deviations and fluctuations, and expectedly followed a generic downward slope. These downward slopes are common for BC trends as the number of discourse units in the community increases and the community's interest fizzles out over time. The decreasing gradients are indicators of the rate at which the content and ideas in the discourse units are losing communal interest, attention, and usage. A gentle gradient means a gradual or negligible loss of sharing and interest by the community in the discourse unit's content and ideas, whereas a much steeper gradient, as indicated by DU8 and DU18 in Figure 3, denotes a more severe decline of interest and usage.

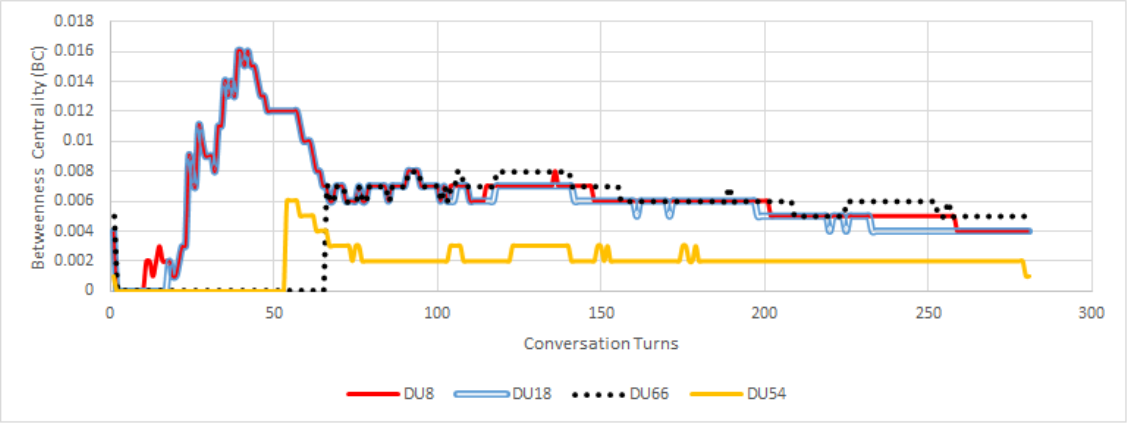

Figure 3. Instructor input (DU54) relative to other participant inputs (DU8, DU18, and DU66) over the period of discourse.

By mid-discourse, both DU8 and DU18 were only able to retain a fair share of communal interest and usage, reflected by the drop in BC values from their initial peaks. This could be due to waning interest in the ideas within the discourse units, an increased pool of ideas contributed by other participants that competes for a limited amount of communal attention, or a combination of these. As for DU66, the discourse unit was introduced later and did not exhibit a similarly high peak, but the BC trend of DU66 still rose to a relatively high level and fluctuated slightly before settling into a slow decline over time, overlapping with the BC trends of the other participant inputs (DU8 and DU18). DU66 represents a discourse unit with some promising ideas even though they were introduced relatively late in the discourse.

In addition to the three chosen discourse units that contain promising ideas, we included instructor inputs (DU 54) to show their relative promisingness when compared to participant inputs (Figure 3). As the keywords were generated based on communal discourse, instructor inputs seem to be of lower interest, but are still somewhat promising as DU54's ideas and content were still shared and discussed among the community. This was indicated by the relatively smaller but continuous BC trend of DU54 over the period of discourse, as compared to the larger BC trends of the other three discourse units containing promising

(2017). Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *Journal of Learning Analytics*, *4*(3), 76–101. http://dx.doi.org/10.18608/jla.2017.43.5

ideas (DU8, DU18, and DU66). The different types of BC trends suggest that the pattern of communal interests in ideas could be predicted using network measures such as BC. We verified these observations by scrutinizing the discourse units qualitatively, and found that the nature and contents of the DUs are consistent with our interpretation of how the community perceived the discourse units.

Other than the discussed BC network measure, we introduced the DC measure and plotted pairs of DC-BC values on a DC-BC graph. A static snapshot at a temporal juncture of the discourse can be represented by a plot of DC-BC values for all discourse units. By considering multiple static snapshots over a period of discourse, we are able to visualize the movement of discourse units on a DC-BC graph. These snapshots can be captured at any conversation turn during discourse, depending on the goals of the analyst. We were interested in finding out changes in the promisingness of the DU's content and ideas as discourse progresses. One possibility was to examine the DC-BC values of DUs between mid-discourse and the end of discourse. We superimposed the DC-BC graphs captured at mid-discourse and at the end of discourse onto a single graph to show the shifting of discourse units over the entire period. The resulting DC-BC graph (Figure 4) reflects the mobility of selected discourse units containing promising ideas among other surrounding discourse units.

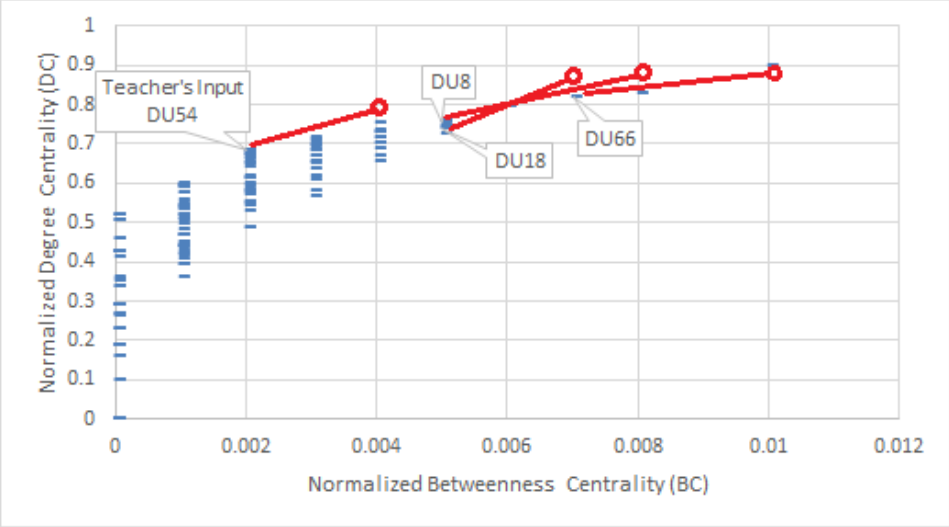

Figure 4. Visualization of moving discourse units containing promising ideas on the superimposed DC-BC graph, from mid-discourse (hollow circle) to the end of discourse (labelled).

From the analysis of the DC-BC graph (Figure 4), discourse units generally moved from the right to the left as discourse progressed, indicating decreasing relevancy to the community and reduced sharing of content and ideas within the discourse units. To further understand how this shift affected the promisingness of ideas in discourse, we used *k*-means clustering as an additional analysis tool to differentiate discourse units from each other and label the types of ideas found in knowledge building discourse.

(2017). Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *Journal of Learning Analytics*, *4*(3), 76–101. http://dx.doi.org/10.18608/jla.2017.43.5

4.3 Recognizing Types of Ideas Using *K*-Means Clustering

The usage of temporal analytics assisted us in identifying promising ideas and discovering idea mobility in discourse using a DC-BC graph. To recognize the types of ideas that emerge from discourse, *k*-means clustering was implemented to suggest groups of discourse units based on the patterns in the data. Following suggested methods in this study, discourse units that possess relatively high DC and BC values are mainly located in the top right quadrant of the DC-BC graph. The starting choice of "k" in this study was three and thus we expect three clusters to be labelled, consisting of groups of discourse units representing promising, potential, and trivial ideas.

4.3.1 Identifying different idea types in discourse

The *k*-means clustering was run several times and the best was chosen to be presented, with reasonably clustered discourse units and minimal overlaps. A snapshot (Figure 5) was taken at the end of discourse, showing the positions of all discourse units plotted on the DC-BC graph and grouped into three clusters.

Results show that the discourse units in cluster 3, represented by the diamonds in Figure 5, consist mostly of discourse units with high DC values and a range of BC values. This means that the ideas in cluster 3 contain a high degree of relevancy and sustain the interest of the community. Discourse units from cluster 2 are represented by the dash symbols and contain ideas of significant DC value, but with a lower range of BC values. The ideas in cluster 2 are, therefore, of lower relevance and interest to the community. Last, discourse units in cluster 1 are represented by the "X" symbol; they possess low DC and low or no BC, indicating that the ideas are barely promising or relevant. These ideas are not used for mediating information, being either isolated instances of irrelevant ideas or just not interesting to the community.

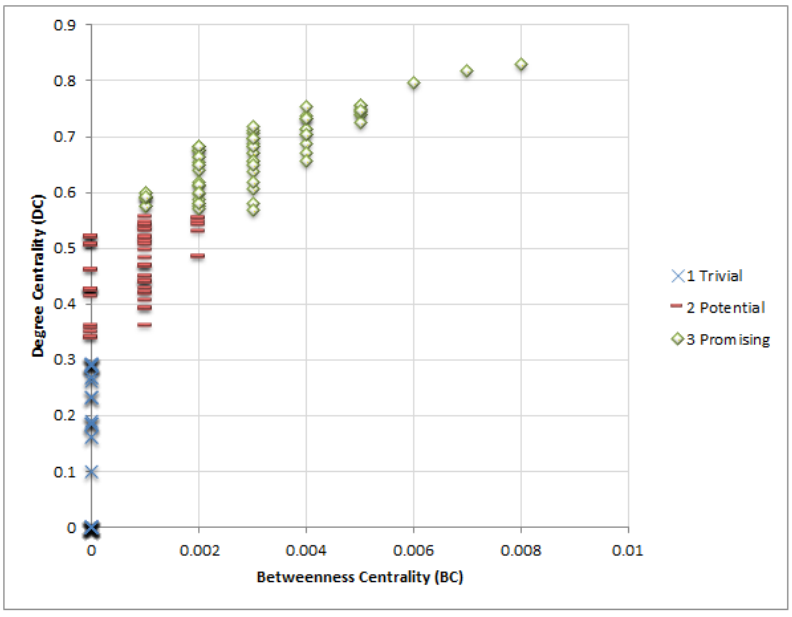

Figure 5. Discourse units positioned in three clusters on the DC-BC graph after *k*-means clustering.

(2017). Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *Journal of Learning Analytics*, *4*(3), 76–101. http://dx.doi.org/10.18608/jla.2017.43.5

With reference to the idea types and respective definitions, we suggest that the discourse units within the three clusters can be labelled as *promising*, *potential*, or *trivial* ideas. Discourse units in cluster 3 are the most promising ideas, with high connection and good mediation capability. Discourse units in cluster 2 represent potential ideas that can be further improved to increase community interest, as the connections to other discourse units are already formed. The remaining discourse units in cluster 1 are trivial ideas that lack relevancy and are mostly uninteresting to the community. The *k*-means algorithm was thus able to group the discourse units into three clusters that represent the different idea types present in the discourse.

4.3.2 Detection of undiscovered promising ideas in discourse

Apart from the discourse units with promising ideas (DU8, DU18, and DU66), identified and verified in previous studies, we discovered two additional discourse units (DU58 and DU271, flagged in Figure 6), using the proposed methods, which were undetected in prior studies.

These two additional distinct discourse units were flagged because they exhibited relatively higher DC and BC values than others. They were not previously uncovered using temporal analytics and BC trends, but rather were discovered after the visualization of discourse units on the DC-BC graph and the application of *k*-means clustering. Qualitative analysis (see Section 4.5) was used to explain the contents in the discourse units and their promisingness of ideas were subsequently verified. A possible explanation may be the introduction of DC as an additional measure, and that the integration of its usage with the proposed DC-BC graph led us to the discovery of additional promising ideas.

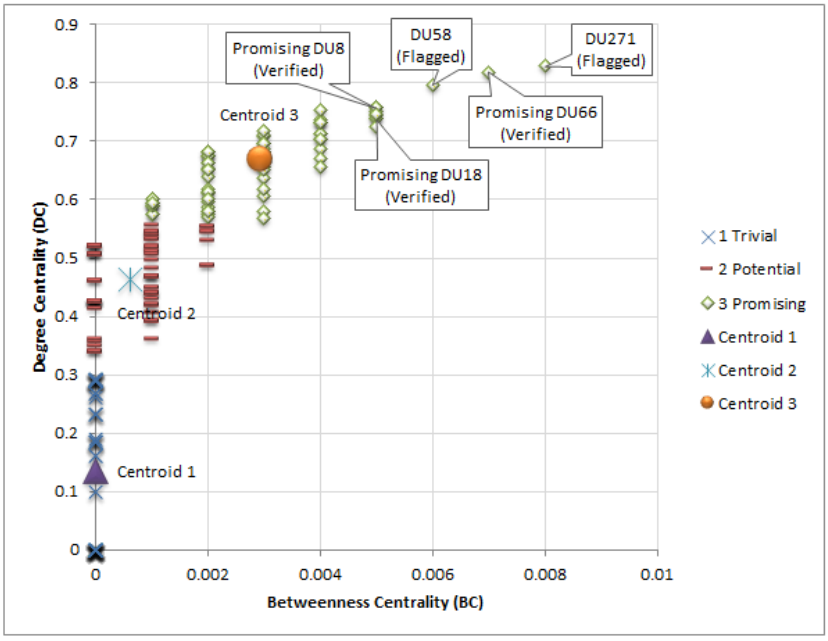

Figure 6. Verified (DU8, DU18, and DU66) and flagged (DU58 and DU271) discourse units containing promising ideas, positioned in the DC-BC graph with cluster centroids.

(2017). Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *Journal of Learning Analytics*, *4*(3), 76–101. http://dx.doi.org/10.18608/jla.2017.43.5

So far, we have shown that *k*-means clustering can be used to group discourse units into clusters with the usage of the DC and BC measures, and more robustly assist in the identification of idea types from discourse units. More importantly, the community can spend more time on ideas promising to the collective advancement of communal knowledge discourse, and not be sidetracked by other facets.

4.4 Role of Instructor Input Among Other Discourse Units

The visualized movement of discourse units about cluster centroids is akin to the representation of ideas that revolve around group-centric positions of idea groups in the whole discourse. In a similar situation, the movement of guiding inputs, such as those from instructors, can be monitored together with cluster centroids representing the cluster of discourse units to reveal how the centre of discussions has shifted over time and how guiding inputs affect learning behaviours among participants. We therefore focused on a promising instructor input (DU54; cluster 3) and tracked it with respect to the cluster centroid 3, which is representative of discussions from other discourse units in the same cluster.

The DC-BC graphs in Figure 7 show the shifting positions of DU54 from mid-discourse to end of discourse. At mid-discourse, DU54 had promising ideas, reflected by the grouping of instructor inputs within cluster 3, but became a borderline case by the end of discourse. The movement of instructor input is a significant lateral shift to the left, as it moved from the right of cluster 3's centroid during discourse, to the left of cluster 3's centroid in a more discourse-centric position at the end of discourse. Although we earlier concluded that instructor input was not as promising as student ideas (Figure 3), we also note that the instructor role in discourse is not solely that of a facilitator and co-creator of knowledge. The instructor has other responsibilities to ensure students who are unable to engage in promising ideas are provided with sufficient assistance and scaffolds to interact and contribute to the community, and continue to share and advance knowledge as a community. Hence, the final position of instructor input on the DC-BC graph is not surprising, as the input became more discourse-centric and was used as a reference by students.

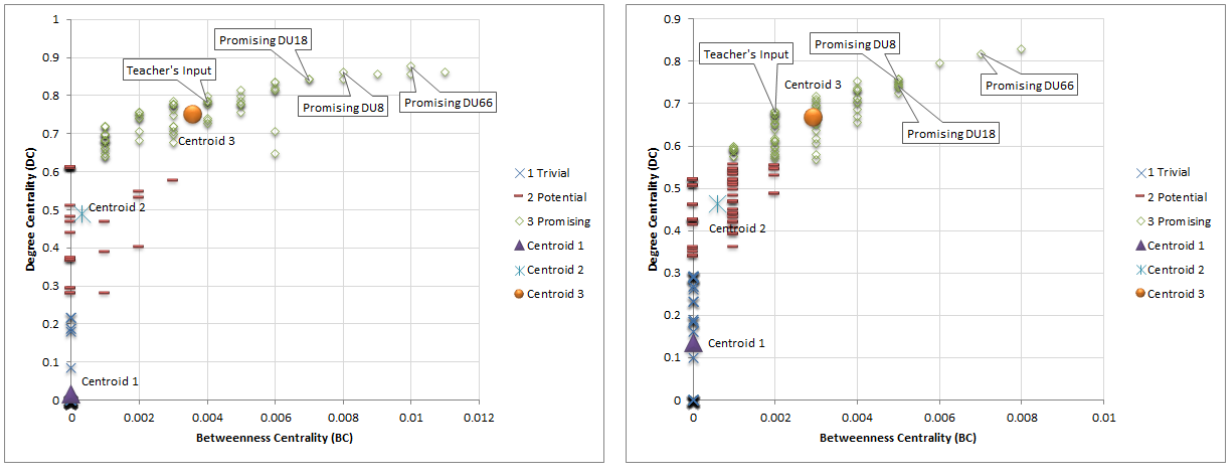

Figure 7. Instructor input at mid-discourse (left) and at the end of discourse (right).

(2017). Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *Journal of Learning Analytics*, *4*(3), 76–101. http://dx.doi.org/10.18608/jla.2017.43.5

4.5 Verification of Discourse Units Using Qualitative Analysis

4.5.1 Verifying promisingness of discourse units using current methods

The dataset used in this paper was previously analyzed using BC trends (Lee & Tan, 2017) and discourse units with promising ideas (DU8, DU18, and DU66) were found. The contents were qualitatively analyzed and verified to be promising ideas. We ascertain our findings by tracking the mentioned discourse units using the DC-BC graph and performing *k*-means clustering, as proposed in this paper. Results from the DC-BC plot visualization and clustering process revealed that the three identified discourse units (DU8, DU18, and DU66) were grouped under cluster 3. These discourse units were also situated in close proximity to the top right quadrant of the DC-BC graph (see Figure 6), which is a region for discourse units with high DC and BC values, thus suggesting that these discourse units contain a strong presence of promising ideas.

4.5.2 Verification of newly discovered discourse units containing promising ideas

The additionally discovered discourse units with promising ideas (DU58 and DU271), detected using the methods described in this paper, were however yet to be verified. Since training data was not used, we further note that instead of performing a two-fold cross validation, we conducted a qualitative analysis to verify the idea types against the qualitative content in the respective discourse units. The following is an excerpt of the qualitative analysis for the two flagged discourse units, DU58 and DU271.

Participant S7 contributed DU58 in the early to midway portion of discourse and the discourse unit contained a list of knowledge building characteristics identified from a conference report of students investigating the topic of cockroaches. Participant S7 listed the knowledge building characteristics observed in the report along with respective evidence, which the participant believes should substantiate the claims reported in the discourse unit:

> 1. Knowledge advancement as a community rather than individual achievement —> cross talk between students to discuss their observations and this sparked further discussions and question from other members of the class (p. 7)
>
> 2. Knowledge advancement as idea improvement rather than as progress toward true or warranted belief —> There is extended discussion on how cockroaches survive ice age (case 2). The students kept throwing forth hypotheses trying to explain this phenomenon.
>
> 3. Knowledge of in contrast to knowledge about —> The teachers don't front load the students with prescribed concepts and explicit skills. They allow the students to generalise and internalise i.e., learn as they do.

The truncated list of characteristics was scrutinized, and participant S7 was able to identify knowledge building characteristics in the report and relate that with what was learnt in the course. During the early stages of discourse, participants including participant S7 were able to point out major characteristics and find supporting evidence with the relevant observations at cited pages or case numbers. The community was then able to focus on understanding these summarized points of information and swiftly gained deeper understanding, building onto this knowledge instead of spending extra time searching through the

(2017). Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *Journal of Learning Analytics*, *4*(3), 76–101. http://dx.doi.org/10.18608/jla.2017.43.5

whole discourse and literature for similar sources of information. We determined that participant S7 has, therefore, contributed a fairly promising quote of relevancy and interest to the community.

The other discourse unit, DU271, was contributed by participant S9 near the end of discourse and contains the following text:

> posted my questions in beginning
>
> 1. what is difference between knowledge building and knowledge creation?
>
> [Questions 2 and 3]
>
> ans: end of our kb lesson, i'm able to answer my own questions.
>
> 1) I understand that knowledge building involves creative, sustained work with ideas around authentic questions and problems, where the overall objective is to work collaboratively to improve those ideas. …
>
> [Answers to questions 2 and 3]
>
> thankful to T1 being able to show what kb is about and as a real example to show us to carry out kb in classroom.

By the end of discourse, participant S9 was able to answer all of the self-initiated inquiries in DU 271 by participating in many of the discussions, reflection and sharing sessions during discourse, and constructing alternative views and opinions leading up to DU271. Participant S9 was, therefore, able to consolidate findings throughout the whole discourse, and reflected on the shared information to produce a final note (DU271) to encapsulate an overall understanding of knowledge building. DU271 is considered a product of the "rise-above" in knowledge building, because participant S9 was able to work with diverse sources and adapt to progressive conditions that constantly require the re-evaluation of one's own knowledge. By constantly improving ideas to achieve new synthesis, S9 was able to move beyond basic understanding and use external resources to back up claims; therefore, DU271 contains promising ideas.

Overall, we were able to verify the presence of promising ideas in selected discourse units, and ascertained that the DC-BC graph can be used with *k*-means clustering to identify promising ideas in discourse. Further, the two additional discourse units detected using the proposed clustering method, were also verified to contain promising ideas through qualitative analysis. The results showed that using the DC-BC graph with clustering can robustly and effectively identify promising ideas in knowledge building discourse.

4.6 Limitations of Findings

We acknowledge some limitations regarding conducting discourse analysis in a post hoc manner, our limitations for the machine learning methods reported, and discuss our options and potential solutions. First, the idea identification and analysis (I^2A) methodology was designed to analyze ideas in discourse

(2017). Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *Journal of Learning Analytics*, *4*(3), 76–101. http://dx.doi.org/10.18608/jla.2017.43.5

between any two temporal junctures and handle both ongoing and completed discourse. A post hoc analysis of discourse data has been presented and discussed in this paper, and we are currently in the midst of analyzing and validating results from the implementation of I^2A in an ongoing discourse. Initial results have shown that I^2A operated similarly and exhibited comparable results for both ongoing and completed discourses, thus showing that the I^2A methodology is generalizable and can be used in a broader scope.

Next, the introduction of machine learning is not novel, but is nonetheless uncommon when applied to educational research. There is no objectively correct clustering algorithm that can be used, and we do not have a mathematical model to justify whether one cluster model should be used over the other. However, we have considered multiple classification, clustering algorithms, and two other factors, namely the ease of implementation and convenience of use, which stand out as important considerations in choosing *k*-means clustering for our discourse data. First, *k*-means clustering does not require appropriate training data to train the clustering algorithm. This would be an additional burden if there were insufficient or incomplete data to use for training purposes. Secondly, *k*-means clustering only requires an input "k," which is the number of expected clusters, to begin the process of clustering. Further, even though we have explained how we can estimate a good "k" value for the clustering process, an alternative method of optimizing the "k" value is to compute the sum of the squared error (SSE) to optimize the value of "k" itself, which we will not elaborate here. In essence, by plotting "k" against SSE, we expect the error of the solution to decrease as "k" is increased, so that an optimized value of "k" can be achieved that is appropriate for the problem without compromising too much on performance and results. Although this process is more elaborate and accurate, in this study, the benefits of finding an accurate final value of k might not be worth the extra processing required by implementing SSE.

Among other machine learning algorithms, *k*-means clustering is also similar to another "nearest neighbour" (*k*-NN) algorithm, which is used for classifying data, as a subset of supervised learning. Both *k*-means and *k*-NN techniques are sensitive to the local structure of the data, and may produce local optimum problems. This means that optimization of discourse analysis may not be eventually achieved, even if the tests are conducted over multiple runs, especially if the starting points for clustering are not wildly varied. In this study, we attempted to cluster discourse units on a DC-BC graph, using unsupervised learning without training sets. We eventually used the *k*-means algorithm and chose the best results out of multiple runs to reduce the local optimum problem.

5 CONCLUSION

In this paper, we introduced the use of temporal analytics with machine learning techniques to investigate ideas promising to the collective advancement of communal knowledge in online knowledge building discourse. We used text-mining procedures to identify conceptual keywords from discourse to construct a discourse unit network. Network measures such as BC and DC were calculated from the discourse unit network, and the temporal analysis of BC trends and visualization of DC-BC values on a two-dimensional graph helped to identify promising ideas and describe idea mobility over time. Following this, *k*-means

(2017). Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *Journal of Learning Analytics*, *4*(3), 76–101. http://dx.doi.org/10.18608/jla.2017.43.5

clustering was conducted, which helped to identify additional promising ideas that were previously undetected using only BC trends. The promising discourse inputs were verified using qualitative analysis, and further demonstrated that promising ideas in discourse can be more robustly found using clustering with the DC-BC graph. The application of the *k*-means clustering algorithm also provided centroids that represent centres of discussion for clusters of discourse units. The movement of ideas, such as instructor input around cluster centroids, show how ideas can affect learning behaviours among the participants. Overall, the consolidated results from the implementation of temporal analytics and the clustering process can provide insights and feedback to users about idea-related processes in discourse, and should better inform users on ideas that are promising to the collective advancement of communal knowledge in knowledge building discourse.

6 ACKNOWLEDGEMENTS

This research study was supported by the Centre for Research and Development in Learning, Nanyang Technological University (CRADLE@NTU). The research team would also like to thank the instructors and in-service teachers who participated in this study.

REFERENCES

Antonenko, P. D., Toy, S., & Niederhauser, D. S. (2012). Using cluster analysis for data mining in educational technology research. *Educational Technology Research and Development*, *60*(3), 383–398. http://dx.doi.org/10.1007/s11423-012-9235-8

Baer, L., & Campbell, J. (2012). From metrics to analytics, reporting to action: Analytics' role in changing the learning environment. In D. G. Oblinger (Ed.), *Game changers: Education and information technologies* (pp. 53–65). EDUCAUSE.

Bailey, K. (1994). Numerical taxonomy and cluster analysis. *Typologies and Taxonomies*, *34*, 24.

Baker, P. (2006). *Using corpora in discourse analysis*. A&C Black.

Bakharia, A., & Dawson, S. (2011). SNAPP: A bird's-eye view of temporal participant interaction. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge* (LAK '11), 27 February–1 March 2011, Banff, AB, Canada (pp. 168–173). New York: ACM. http://dx.doi.org/10.1145/2090116.2090144

Bereiter, C., & Scardamalia, M. (1993). *Surpassing ourselves: An inquiry into the nature and implications of expertise.* Chicago: Open Court.

Chen, B. (2014). Promisingness judgements *as facilitators of knowledge building in elementary science* (Doctoral dissertation). University of Toronto, Toronto, Ontario, Canada.

Chen, B. (2017). Fostering scientific understanding and epistemic beliefs through judgments of promisingness. *Educational Technology Research and Development, 65*(2), 255–277. http://dx.doi.org/10.1007/s11423-016-9467-0

Chen, B., Haklev, S., Harrison, L., Najafi, H., & Rolheiser, C. (2015). How do MOOC learners' intentions relate to their behaviors and overall outcomes? Poster presented at the American Educational Research Association Annual Conference (AERA 2015), 16–20 April 2015, Chicago, IL, USA.

(2017). Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *Journal of Learning Analytics*, *4*(3), 76–101. http://dx.doi.org/10.18608/jla.2017.43.5

Chen, B., Scardamalia, M., & Bereiter, C. (2015). Advancing knowledge-building discourse through judgments of promising ideas. *International Journal of Computer-Supported Collaborative Learning*, *10*(4), 345–366. http://dx.doi.org/10.1007/s11412-015-9225-z

Chen, B., Wise, A. F., Knight, S., & Cheng, B. H. (2016). Putting temporal analytics into practice: The 5th International Workshop on Temporality in Learning Data. *Proceedings of the 6th International Conference on Learning Analytics and Knowledge* (LAK '16), 25–29 April 2016, Edinburgh, UK (pp. 488–489). New York: ACM. http://dx.doi.org/10.1145/2883851.2883865

Chiu, M. M., & Fujita, N. (2014a). Statistical discourse analysis of online discussions: Informal cognition, social metacognition, and knowledge creation. In *Knowledge Creation in Education* (pp. 97–112). Springer Singapore. http://dx.doi.org/10.1007/978-981-287-047-6_6

Chiu, M. M., & Fujita, N. (2014b). Statistical discourse analysis: A method for modeling online discussion processes. *Journal of Learning Analytics*, *1*(3), 61–83. http://dx.doi.org/10.18608/jla.2014.13.5

Estivill-Castro, V. (2002). Why so many clustering algorithms: A position paper. *ACM SIGKDD Explorations Newsletter*, *4*(1), 65–75. http://dx.doi.org/10.1145/568574.568575

Finley, T., & Joachims, T. (2005). Supervised clustering with support vector machines. *Proceedings of the 22nd International Conference on Machine Learning* (ICML-05), 07–11 August 2005, Bonn, Germany (pp. 217–224). New York: ACM. http://dx.doi.org/10.1145/1102351.1102379

García, N. A., & Caplan, A. (2014). Reading the world's classics critically: A keyword-based approach to literary analysis in foreign language studies. *Critical Inquiry in Language Studies*, *11*(2), 100–120.

Gardner, H. (1994). More on private intuitions and public symbol systems. *Creativity Research Journal*, *7*(3–4), 265–275. http://dx.doi.org/10.1080/15427587.2014.906801

Garrard, P., Rentoumi, V., Gesierich, B., Miller, B., & Gorno-Tempini, M. L. (2014). Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse. *Cortex*, *55*, 122–129. http://dx.doi.org/10.1016/j.cortex.2013.05.008

Gruzd, A., Paulin, D., & Haythornthwaite, C. (2016). Analyzing social media and learning through content and social network analysis: A faceted methodological approach. *Journal of Learning Analytics*, *3*(3), 46–71. http://dx.doi.org/10.18608/jla.2016.33.4

Hsiao, I. H., & Awasthi, P. (2015). Topic facet modeling: Semantic visual analytics for online discussion forums. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 231–235). New York: ACM. http://dx.doi.org/10.1145/2723576.2723613

Knight, S., & Littleton, K. (2015). Discourse-centric learning analytics: Mapping the terrain. *Journal of Learning Analytics*, *2*(1), 185–209. http://dx.doi.org/10.18608/jla.2015.21.9

Knight, S., Wise, A. F., Chen, B., & Cheng, B. H. (2015). It's about time: 4th International Workshop on Temporal Analyses of Learning Data. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 388–389). New York: ACM. http://dx.doi.org/10.1145/2723576.2723638

Kovanović, V., Joksimović, S., Waters, Z., Gašević, D., Kitto, K., Hatala, M., & Siemens, G. (2016). Towards automated content analysis of discussion transcripts: A cognitive presence case. *Proceedings of*

(2017). Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *Journal of Learning Analytics*, *4*(3), 76–101. http://dx.doi.org/10.18608/jla.2017.43.5

the 6th International Conference on Learning Analytics and Knowledge (LAK '16), 25–29 April 2016, Edinburgh, UK (pp. 15–24). New York: ACM. http://dx.doi.org/10.1145/2883851.2883950

Lakatos, I. (1970). The methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91–195). Cambridge University Press.

Lee, A. V. Y., & Tan, S. C. (2017). Temporal analytics with discourse analysis: Tracing ideas and impact on communal discourse. *Proceedings of the 7th International Conference on Learning Analytics and Knowledge* (LAK '17), 13–17 March 2017, Vancouver, BC, Canada (pp. 120–127). New York: ACM. http://dx.doi.org/10.1145/3027385.3027386

Lee, A. V. Y., Tan, S. C., & Chee, K. J. K. (2016). Idea identification and analysis (I^2A): A search for sustainable promising ideas within knowledge-building discourse. In C. K. Looi, J. Polman, U. Cress, & P. Reimann (Eds.), *Transforming Learning, Empowering Learners: Proceedings of the 12th International Conference of the Learning Sciences* (ICLS '16), 20–24 June 2016, Singapore (Vol. 1, pp. 90–97). International Society of the Learning Sciences.

Locke, J. (1836). *An essay concerning human understanding*. London: Balne, Printer, Gracechurch Street.

Martin, R. L. (2009). *The design of business: Why design thinking is the next competitive advantage*. Boston, MA: Harvard Business Press. http://dx.doi.org/10.1111/j.1948-7169.2010.00039.x

Mercer, N. (2008). The seeds of time: Why classroom dialogue needs a temporal analysis. *The Journal of the Learning Sciences*, *17*(1), 33–59. http://dx.doi.org/10.1080/10508400701793182

Merriam-Webster (n.d.). Idea [Def. 1a]. Retrieved 2 Nov. 2016 from http://www.merriam-webster.com/dictionary/idea.

Molenaar, I. (2014). Advances in temporal analysis in learning and instruction. *Frontline Learning Research*, *2*(4), 15–24. http://dx.doi.org/10.14786/flr.v2i4.118

Nayyar, Z., Hashmi, N., Rafique, N., & Mahmood, K. (2016). Keyword based searching in social networks. *Proceedings of the SAI Computing Conference* (SAI 2016), 13–15 July 2016, London, United Kingdom (pp. 701–705). IEEE Computing Society. http://dx.doi.org/10.1109/SAI.2016.7556058

Oshima, J., Oshima, R., & Fujita, W. (2016). Refinement of semantic network analysis for epistemic agency in collaboration. In C. K. Looi, J. Polman, U. Cress, & P. Reimann (Eds.), *Transforming Learning, Empowering Learners: Proceedings of the 12th International Conference of the Learning Sciences* (ICLS '16), 20–24 June 2016, Singapore (Vol. 2, pp. 1191–1192). International Society of the Learning Sciences.

Oshima, J., Oshima, R., & Matsuzawa, Y. (2012). Knowledge building discourse explorer: A social network analysis application for knowledge building discourse. *Educational Technology Research and Development*, *60*(5), 903–921. http://dx.doi.org/10.1007/s11423-012-9265-2

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., & Smith, N. A. (2013). *Improved part-of-speech tagging for online conversational text with word clusters.* Association for Computational Linguistics.

Paavola, S., & Hakkarainen, K. (2005). The knowledge creation metaphor: An emergent epistemological approach to learning. *Science & Education, 14,* 535–557. http://dx.doi.org/10.1007/s11191-004-5157-0

(2017). Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *Journal of Learning Analytics*, *4*(3), 76–101. http://dx.doi.org/10.18608/jla.2017.43.5

Piety, P. J., Hickey, D. T., & Bishop, M. J. (2014). Educational data sciences: Framing emergent practices for analytics of learning, organizations, and systems. *Proceedings of the 4th International Conference on Learning Analytics and Knowledge* (LAK '14), 24–28 March 2014, Indianapolis, IN, USA (pp. 193–202). New York: ACM. http://dx.doi.org/10.1145/2567574.2567582

Reategui, E., Epstein, D., Lorenzatti, A., & Klemann, M. (2011). Sobek: A text mining tool for educational applications. In R. Stahlbock (Ed.), *Proceedings of the 7th International Conference on Data Mining* (DMIN '11), 18–21 July 2011, Las Vegas, Nevada, USA (pp. 59–64). http://worldcomp-proceedings.com/proc/p2011/DMI2159.pdf

Reimann, P. (2009). Time is precious: Variable- and event-centred approaches to process analysis in CSCL research. *International Journal of Computer-Supported Collaborative Learning*, *4*(3), 239–257. http://dx.doi.org/10.1007/s11412-009-9070-z

Rosé, C., Wang, Y. C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., & Fischer, F. (2008). Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning*, *3*(3), 237–271. http://dx.doi.org/10.1007/s11412-007-9034-0

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 210–229. http://dx.doi.org/10.1147/rd.441.0206

Scardamalia, M. (2002). Collective cognitive responsibility for the advancement of knowledge. In B. Smith (Ed.), *Liberal Education in a Knowledge Society*, *97*, 67–98.

Scardamalia, M. (2004). CSILE/Knowledge forum®. *Education and technology: An encyclopedia*, 183–192. Santa Barbara, CA: ABC-CLIO.

Scardamalia, M., & Bereiter, C. (2003). Knowledge building. *Encyclopedia of education* (pp. 1370–1373). New York: Macmillan Reference.

Scardamalia, M., & Bereiter, C. (2006). Knowledge building: Theory, pedagogy, and technology. In R. K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 97–115). New York: Cambridge University Press.

Schenker, A. (2003). *Graph-theoretic techniques for web content mining* (Doctoral dissertation). University of South Florida, Tampa, Florida, USA.

Schoor, C., & Bannert, M. (2012). Exploring regulatory processes during a computer-supported collaborative learning task using process mining. *Computers in Human Behavior*, *28*(4), 1321–1331. http://dx.doi.org/10.1016/j.chb.2012.02.016

Sfard, A. (1998). On two metaphors for learning and the dangers of choosing just one. *Educational Researcher, 27*(2)*,* 4–13.

Sun, W., Zhang, J., Jin, H., & Lyu, S. (2014). Analyzing online knowledge-building discourse using probabilistic topic models. In J. L. Polman, E. A. Kyza, D. K. O'Neill, I. Tabak, W. R. Penuel, A. S. Jurow, & L. D'Amico (Eds.), *Learning and Becoming in Practice: Proceedings of the International Conference of the Learning Sciences* (ICLS '14), 23–27 June 2014, Boulder, CO, USA (Vol. 2, pp. 823–830). International Society of the Learning Sciences.

(2017). Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *Journal of Learning Analytics*, *4*(3), 76–101. http://dx.doi.org/10.18608/jla.2017.43.5

Suthers, D. D. (2006). Technology affordances for intersubjective meaning making: A research agenda for CSCL. *International Journal of Computer-Supported Collaborative Learning*, *1*(3), 315–337. http://dx.doi.org/10.1007/s11412-006-9660-y

Wilkowski, J., Deutsch, A., & Russell, D. M. (2014). Student skill and goal achievement in the mapping with google MOOC. *Proceedings of the 1st ACM Conference on Learning @ Scale* (L@S 2014), 4–5 March 2014, Atlanta, Georgia, USA (pp. 3–10). New York: ACM. http://dx.doi.org/10.1145/2556325.2566240

Zhang, J., Chen, M. H., Tao, D., Naqvi, S., & Peebles, B. (2014). Using idea thread mapper to support collaborative reflection for sustained knowledge building. Poster presented at the American Educational Research Association Annual Conference (AERA 2014), 3–7 April 2014, Philadelphia, PA, USA. Retrieved from https://tccl.arcc.albany.edu/wpsite/wp-content/uploads/Final-Version-ITM-2year_PosterPaper.pdf

Zhang, J., Scardamalia, M., Reeve, R., & Messina, R. (2009). Designs for collective cognitive responsibility in knowledge-building communities. *The Journal of the Learning Sciences*, *18*(1), 7–44. http://dx.doi.org/10.1080/10508400802581676