Chapter 7: Content Analytics: The Definition, Scope, and an Overview of Published Research

Vitomir Kovanović[1], Srećko Joksimović[2], Dragan Gašević[1,2], Marek Hatala[3], George Siemens[4]

[1] School of Informatics, The University of Edinburgh, United Kingdom
[2] Moray House School of Education, The University of Edinburgh, United Kingdom
[3] School of Interactive Arts and Technology, Simon Fraser University, Canada
[4] LINK Research Lab, The University of Texas at Arlington, USA

DOI: 10.18608/hla17.007

ABSTRACT

The field of learning analytics recently attracted attention from educational practitioners and researchers interested in the use of large amounts of learning data for understanding learning processes and improving learning and teaching practices. In this chapter, we introduce *content analytics* – a particular form of learning analytics focused on the analysis of different forms of educational content. We provide the definition and scope of content analytics and a comprehensive summary of the significant content analytics studies in the published literature to date. Given the early stage of the learning analytics field, the focus of this chapter is on the key problems and challenges for which existing content analytics approaches are suitable and have been successfully used in the past. We also reflect on the current trends in content analytics and their position within a broader domain of educational research.

Keywords: Content analytics, learning content

With the large amounts of data related to student learning being collected by digital systems, the potential for using this data for improving learning processes and teaching practices is widely recognized (Gašević, Dawson, & Siemens, 2015). The emerging field of learning analytics recently gained significant attention from educational researchers, practitioners, administrators, and others interested in the intersection of technology and education and the use of this vast amount of data for improving learning and teaching (Buckingham Shum & Ferguson, 2012). Among the different types of data, the analysis of learning content is commonly used for the development of learning analytics systems (Buckingham Shum & Ferguson, 2012; Chatti, Dyckhoff, Schroeder, & Thüs, 2012; Ferguson, 2012; Ferguson & Buckingham Shum, 2012). These include various forms of data produced by instructors (course syllabi, documents, lecture recordings), publishers (textbooks), or students (essays, discussion messages, social media postings). In this chapter, we introduce *content analytics*, an umbrella term used to refer to different types of learning analytics focusing on the analysis of various forms of learning content. We further provide a critical reflection on the state of the content analytics domain, identifying potential shortcomings and directions for future studies. We begin by discussing different forms of learning content and commonly adopted definitions of content analytics. Special attention is given to the range of problems commonly addressed by content analytics, as well as to various methodological approaches, tools, and techniques.

Learning Content and Content Analytics

According to Moore (1989), the defining characteristic of any form of education is the interaction between learners and learning content. Without content "there cannot be education since it is the process of intellectually interacting with the content that results in changes in the learner's understanding, the learner's perspective, or the cognitive structures of the learner's mind" (p. 2). While the most commonly used forms of educational content are written materials (Cook, Garside, Levinson, Dupras, & Montori, 2010), the ubiquitous access to personal computers and the Internet resulted in both a broad availability of learning resources and increased use of interactive and multimedia educational resources. Likewise, the emergence of web-based systems such as blogs and online discussion forums, and popular social media platforms (Twitter, Facebook) introduced a new dimension and provided access to a relatively new set of learner-generated resources (De Freitas, 2007, p. 16). The overall result is that landscape of educational content is expanding and diversifying, bringing along a new set of potential advantages, benefits, challenges, and risks (De Freitas, 2007). This global trend also creates fertile ground for the development of novel learning analytics approaches.

To provide an overview of content analytics literature, we should first define what is meant by content analytics. We define content analytics as

> Automated methods for examining, evaluating, indexing, filtering, recommending, and visualizing different forms of digital learning content, regardless of its producer (e.g., instructor, student) with the goal of understanding learning activities and improving educational practice and research.

This definition reveals that content analytics focuses on the automated analysis of the different "resources" (textbooks, web resources) and "products" (assignments, discussion messages) of learning. This is in clear contrast to analytics focused on the analysis of student behavioural data, such as the analysis of trace data from learning management systems. Although in general students can produce learning content of different types (text, video, audio), given the present state of educational technologies, and online/blended learning pedagogies, the content produced by the learners is predominantly text-based (assignment responses, discussion messages, essays). While there are cases where students produce non-textual content (video recordings of their presentations), they still represent a relative minority; consequently, very few analytical systems have been developed. Thus, the focus of this chapter is predominantly on text-based learning content, despite the broader definition of content analytics, which also encompasses multimedia learning content.

We should point out that content analytics is primarily defined in terms of the application domain, as many of the tools and techniques used are also employed in other types of learning analytics. As such, content analytics encompasses several more specific forms of analytics, including discourse analytics (Knight & Littleton, 2015), writing analytics (Buckingham Shum et al., 2016), assessment analytics (Ellis, 2013), and social learning analytics (Buckingham Shum & Ferguson, 2012). These particular analytics define their foci more specifically to examine learning content produced in particular learning products, processes, or contexts. As a consequence, our definition is broader than, for example, the definition of social content analytics by Buckingham Shum and Ferguson (2012), as a "variety of automated methods that can be used to examine, index and filter online media assets, with the intention of guiding learners through the ocean of potential resources available to them" (p. 15). We argue that the definition of content analytics used in this report – which does not focus on a particular learning setting or process – enables the development of standard analytical approaches applicable to many similar learning domains. Given the early stage of learning analytics development, the focus on the type of learning materials and the methodologies, techniques, and tools for their analysis promotes the establishment of community-wide standards of conducting content analytics research, which is critical for the advancement of the learning analytics field.

It is important to emphasize the difference between *content analysis* (Krippendorff, 2003) and content analytics, which are both commonly used techniques in educational research (Ferguson & Buckingham Shum, 2012). Despite similar names, content analysis is a much older and well-established research technique widely used across social sciences, including research in education, educational technology, and distance/online education (De Wever, Schellens, Valcke, & Van Keer, 2006; Donnelly & Gardner, 2011; Strijbos, Martens, Prins, & Jochems, 2006) to assess latent variables of written text. Given that many of the learning analytics systems are also focused on the examination of latent constructs, a large part of content analytics is an application of computational techniques for the purpose of content analysis (Kovanović, Joksimović, Gašević, & Hatala, 2014). However, content analytics includes different additional forms of analysis, which are not the focus of content analysis, such as assessment of student writings, automated student grading, or topic discovery in the document corpora.

CONTENT ANALYTICS TASKS AND TECHNIQUES

To provide an overview of content analytics, we conducted a review of the published literature on learning analytics and educational technology to identify research studies that made use of content analytics. We looked at the proceedings of the Learning Analytics and Knowledge Conference, the *Journal of Learning*

Analytics, the *Journal of Educational Data Mining*, the *Journal of Artificial Intelligence in Education*, and Google Scholar. After obtaining the relevant studies, we grouped them based on the research problems being addressed. We identified three groups of studies roughly focused on the three main types of data used for content analytics (i.e., learning resources, students' learning products, and students' social interactions). The remainder of this section provides a detailed overview of the identified groups of studies and associated tools and techniques.

Content Analytics of Learning Resources

One of the earliest uses of content analytics was for the analysis of educational resources and materials, and recommendation, organization, and evaluation of those resources. Given the vast amounts of learning materials available to students, one domain of particular interest is the recommendation of relevant learning-related content, based on various criteria such as student interest or course progress (Manouselis, Drachsler, Vuorikari, Hummel, & Koper, 2011; Romero & Ventura, 2010). The development of content analytics systems is typically based on recommender systems technologies, which can be split into two broad categories (Drachsler, Hummel, & Koper, 2008):

1. **Collaborative filtering** (CF) techniques, in which resources being recommended to a student were found by looking for either 1) *related students* (i.e., user-based CF), or 2) *related resources* (i.e., item-based CF). In the former case, students with a substantial overlap in their use of resources probably share common interests; in the latter case, resources used together by a large number of users are likely to be similar.
2. **Content-based** techniques, in which recommendations are discovered by directly comparing the content of resources to be recommended and by looking for most similar resources to the ones a student is currently using or that match the student's profile data.

Both approaches have been extensively used in educational technology (for an overview see Drachsler et al. 2008; Manouselis et al., 2011). For example, Walker, Recker, Lawless, and Wiley (2004) built AlteredVista, a collaborative system for discovering useful educational resources, while Zaldivar, García, Burgos, Kloos, and Pardo (2011) used content-based techniques to recommend course notes to students, based on their document browsing patterns. Content-based methods have also been used to recommend solutions (Hosseini & Brusilovsky, 2014) and relevant examples (Muldner & Conati, 2010) to programming tasks, and even to recommend suitable academic courses (Bramucci & Gaston, 2012). It should also be noted that the quality of recommendations is often dependent on the selection of particular document similarity measures (Verbert et al., 2012), which must be chosen to match the given learning context or activity.

Another important domain represents the automatic organization and classification of different instructional materials (often different learning objects), using automated techniques for keyword extraction, tagging, and clustering. For example, Bosnić, Verbert, and Duval (2010) compared different techniques for keyword extraction from learning objects, while Cardinaels, Meire, and Duval (2005) showed that an analysis of document content, usage, and context could be used to automatically create relevant metadata information for a given learning object. Techniques such as text clustering (Niemann et al., 2012), neural network classifiers (Roy, Sarkar, & Ghose, 2008), and collaborative tagging (Bateman, Brooks, McCalla, & Brusilovsky, 2007) have been used successfully to group, organize, and annotate different learning objects. More recently, with increased use of multimedia in education, different content analytics techniques have been used to automatically find important moments in lecture recordings to enhance navigation and use of video resources (Brooks, Amundson, & Greer, 2009; Brooks, Johnston, Thompson, & Greer, 2013).

In addition to organization and recommendation of learning resources, content analytics has been used to assess the quality of available instructional materials and how they impact learning outcomes. Dufty, Graesser, Louwerse, and McNamara (2006) showed that cohesiveness of the written text, as calculated by the Coh-metrix tool (Graesser, McNamara, & Kulikowich, 2011; McNamara, Graesser, McCarthy, & Cai, 2014), can successfully be used to evaluate the grade-level of textbooks, giving significantly better results than the simple text readability measures (e.g., Flesch Reading Ease, Flesch–Kincaid Grade Level, Degrees of Reading Power). Research has also revealed the direct link between the coherence of the provided learning materials and student comprehension of the subject domain (McNamara, Kintsch, Songer, & Kintsch, 1996; Varner, Jackson, Snow, & McNamara, 2013). The relationship between coherence and comprehension is also moderated by the students' level of background knowledge (Wolfe et al., 1998), which should be taken into account for recommending learning materials.

Content Analytics of Students' Products of Learning

One of the core goals of learning analytics is to enable provision of timely and relevant feedback to learners while studying (Siemens et al., 2011). One of the earliest domains where content analytics has been applied is the analysis of student essays, also known as automated essay scoring (AES). The most widely applied technique for automated essay scoring is Latent Semantic Analysis (LSA) (Landauer, Foltz, & Laham, 1998), used to measure the semantic similarity between two bodies of text through the analysis of their word co-occurrences. In the case of AES, LSA similarity is used to calculate the resemblance of an essay to a predefined set of essays and, based on those similarities, calculate a single, numeric measure of essay quality. In addition to LSA-based measures of essay quality, more recent systems such as WriteToLearn (Foltz & Rosenstein, 2015) include an extensive set of visualizations to provide students with feedback designed to help them acquire essay writing skills. While AES systems have been primarily used for the provision of real-time feedback (Crossley, Allen, Snow, & McNamara, 2015; Foltz et al., 1999; Foltz & Rosenstein, 2015), they could also be used for the (partial) automation of essay grading (Foltz et al., 1999), as they have shown to be as reliable and consistent as human graders.

Besides calculating the similarity of a text to a predefined collection of documents, LSA can also be used for calculating *internal* document similarity, often referred to as document coherence (the more coherent the document, the more semantically similar are its sentences). LSA is the underlying principle behind the Coh-metrix tool (Graesser et al., 2011; McNamara et al., 2014), often used to measure the quality of document writing. Coh-metrix has been extensively utilized for the analysis of different forms of written materials, including essays, discussion messages, and textbooks (McNamara et al., 2014). For example, it was adopted in Writing-Pal (McNamara et al., 2012), which is an intelligent tutoring system that provides students with feedback during essay writing exercises, looking at the essay's cohesiveness (calculated by Coh-metrix) as the indicator of its quality.

Another commonly adopted technique for the assessment of student essays are graph-based visualization methods, also based on a text's word co-occurrences. In addition to assessing the quality of writing, these tools are also used for summarizing essay content. For example, the OpenEssayist system (Whitelock, Field, Pulman, Richardson, & Van Labeke, 2014; Whitelock, Twiner, Richardson, Field, & Pulman, 2015) provides a graph-based overview of a student's essay in order to help the student visualize the relationship between different parts of the essay with the goal of teaching students how to write high-quality essays with a solid structure and a coherent narrative. Graph-based methods are also adopted for automated extraction of concept maps from students' collaborative writings. Such concept maps are then used to provide visual feedback to learners (Hecking & Hoppe, 2015) as a means of helping them review and revise their essays.

Besides approaches based on word co-occurrences, natural language processing techniques have also been used, in particular for the linguistic and rhetorical analysis of student essays. For instance, XIP Dashboard (Simsek, Buckingham Shum, De Liddo, Ferguson, & Sándor, 2014; Simsek, Buckingham Shum, Sandor, De Liddo, & Ferguson, 2013) visualizes meta-discourse of essays and highlights rhetorical moves and functions that help assess the quality of an argument in the essay (Simsek et al., 2014). These approaches to content analytics are also very similar to discourse-centric learning analytics (Buckingham Shum et al., 2013; Knight & Littleton, 2015) given that they use the same set of techniques for understanding the linguistic functions of the different parts of written text.

In addition to analyzing student essays, similar content analytics methods have been used for other types of student writing, most notably short answers (Burrows, Gurevych, & Stein, 2014). In the context of teaching physics, Dzikovska, Steinhauser, Farrow, Moore, and Campbell (2014) built a novel adaptive feedback system that takes into account the content of students' short answers, thus providing contextually relevant feedback. Likewise, the WriteEval system (Leeman-Munk, Wiebe, & Lester, 2014) evaluates students' short answers and provides feedback with follow-up instructions and tasks. As with essay grading, a set of reference answers facilitates the work of this group of systems. Similar approaches are also used for teaching troubleshooting skills (Di Eugenio, Fossati, Haller, Yu, & Glass, 2008), logic (Stamper, Barnes, & Croy, 2010), and PHP programming (Weragama & Reye, 2014). There have also been studies (Ramachandran, Cheng, & Foltz, 2015; Ramachandran & Foltz, 2015) showing the potential of using graph-based techniques for automated discovery of reference answers.

We should also note that many of the content analytics feedback systems have specifically been designed to provide instructors with feedback on student learning activities. For example, Lárusson and White (2012) used visualizations of student essays to inform instructors about the originality in student writings and particular points in time when students start to develop critical thinking. Besides providing feedback to students, automatic extraction of concept maps from student essays was also used to provide instructors with a broad overview of student learning activities (Pérez-Marín & Pascual-Nieto, 2010). Extraction of concept maps was also used for analysis of student chat logs (Rosen, Miagkikh, & Suthers, 2011), which are then used to provide instructors with an overview of social interactions and knowledge building among groups of students. Similarly, types of feedback and their effects on student engagement have also been explored. For instance, Crossley, Varner, Roscoe, and

McNamara (2013) investigated which types of feedback result in the biggest improvement in quality of student writing (based on the Coh-metrix analysis of student essays) while Calvo, Aditomo, Southavilay, and Yacef (2012) investigated how different types of feedback (i.e., directive, reflective) affect student essay editing behaviour. The ways in which students view and annotate video recordings has also been investigated (Gašević, Mirriahi, & Dawson, 2014; Mirriahi & Dawson, 2013) showing the potential for combining the analysis of different types of learning content.

A large body of work has also examined the association between different qualities of student essays and performance. The primary goal of these studies is to understand what encompasses successful writing (Allen, Snow, & McNamara, 2014; Crossley, Roscoe, & McNamara, 2014; McNamara, Crossley, & McCarthy, 2009; Snow, Allen, Jacovina, Perret, & McNamara, 2015), and how it relates to course performance (Robinson, Navea, & Ickes, 2013; Simsek et al., 2015). Current research has also revealed direct links between the coherence of the provided learning materials and the quality of students' reading summaries (Allen, Snow, & McNamara, 2015). Studies have also shown that insights into student comprehension of reading materials can be obtained through the analysis of their reading summaries using Coh-metrix cohesiveness measures and Information Content – a measure of text informativeness (Mintz, Stefanescu, Feng, D'Mello, & Graesser, 2014). Content analytics has also been used for understanding collaborative writing processes by using techniques such as Hidden Markov Models (Southavilay, Yacef, & Calvo, 2009, 2010) and probabilistic topic modelling (e.g., LDA; Southavilay, Yacef, Reimann, & Calvo, 2013). The same techniques are applied to understand how students learn to program (Blikstein, 2011), and even to analyze transcripts of student interviews to assess their expertise (Worsley & Blikstein, 2011) and knowledge of a given domain (Sherin, 2012).

Content Analytics of Students' Social Interactions

In online and distance education, asynchronous online discussions represent one of the primary means of interaction among students, and between students and instructors (Anderson & Dron, 2012). As such, insights into the overall discussion activity and contributions of different students are two areas where content analytics have been successfully applied, often using methods similar to those used for analyzing learning materials (e.g., LSA, Coh-metrix). Using LSA and Social Network Analysis (SNA), Teplovs, Fujita, and Vatrapu (2011) developed a visual analytics system that provides students with an overview of student contributions to online discourse. In addition to SNA, Hever et al. (2007) have also used process mining in combination with content analytics to raise awareness and enable better moderation of online discussions. Through the classification of student discussion messages based on their contribution type, textual content, and relationships (i.e., links) Hever et al. (2007) developed a message classification system that can be used to label discussion messages based on predefined theoretical or pedagogical categories. In addition to online discussions, raising instructor awareness of student activities in social media is explored by the LARAe system (Charleer, Santos, Klerkx, & Duval, 2014) showing the huge potential of social media for understanding student activities and learning progress. LARAe can automatically gather student social media postings (using RSS and Twitter API technologies) and then automatically assign one of 51 different badges to students, based on the observed social media activity. Instructors are then shown the collected information in the form of a dashboard for an easy overview of student activity and its change over time.

Online discussions have also been the focus of education researchers, who typically have used manual content analysis methods for parsing student discussion messages. Over the years, several content analytics systems have been developed to automate this process, in particular, analysis using the popular Community of Inquiry (CoI) framework (Garrison, Anderson, & Archer, 2001). For example, McKlin, Harmon, Evans, and Jones (2002) and McKlin (2004) developed a neural network classification system to automate coding of discussion messages for level of *cognitive presence*, the central construct of the CoI framework, focused on the development of students' critical and deep thinking skills. Building on results by McKlin (2004), a Bayesian network classification is used by the Automated Content Analysis Tool (Corich, Hunt, & Hunt, 2012) to provide a more generalizable model of classification that can be adopted for a wider range of coding schemes besides cognitive presence. More recently, several studies (Kovanović et al., 2014, 2016; Waters, 2015) examined the use of different text-mining techniques for coding messages for level of cognitive presence. Kovanović et al. (2014) developed a support vector machine classifier using different surface-level classification features (i.e., n-grams, part-of-speech n-grams, linguistic dependency triplets, the number of mentioned concepts, and discussion position metrics), which achieved higher classification accuracy than previous reports (McKlin, 2004; McKlin et al., 2002). The study by Waters (2015) also showed the benefits of using the structure of online discussions for text classification using conditional random fields, a structured classification technique that takes into the account relationships (i.e., reply-to structure) among individual classification instances (i.e., discussion messages).

Finally, a study by Kovanović et al. (2016) showed that metrics provided by the Coh-metrix (Graesser et al., 2011) and Linguistic Inquiry and Word Count (LIWC) tools (Tausczik & Pennebaker, 2010) – in combination with some of the NLP and discussion-position features – can be successfully used to develop a classification system almost as accurate as human coders. While further improvements are needed before this system can be widely adopted by educational researchers, the progress is promising and has the potential to advance research practices in content analysis.

With the social-constructivist view of learning and knowledge creation, a large body of work has utilized content analytics for understanding the role of social interactions on knowledge construction. For example, there has been significant research on linguistic differences – as captured by LIWC metrics – in discussion contributions (Joksimović, Gašević, Kovanović, Adesope, & Hatala, 2014; Xu, Murray, Park Woolf, & Smith, 2013) and how those differences relate to student grades (Yoo & Kim, 2012). Similarly, Chiu and Fujita (2014a, 2014b), investigated interdependencies between different types of discussion contributions with statistical discourse analysis (SDA), a group of statistical methods used to provide realistic modelling of student discourse interactions, while Yang, Wen, and Rosé (2014) used LDA and mixed membership stochastic blockmodels (MMSB) to examine what types of student discussion contributions are likely to receive response. Finally, using simple word frequency analysis, Cui and Wise (2015) examined what kinds of contributions are most likely to be acknowledged and answered by instructors. These and similar studies have the goal of understanding how interactions in online discourse eventually shape the learning outcomes and knowledge building. Similarly, different content analytics methods (text classification, topic modelling, mixed membership stochastic blockmodels) and tools (Coh-metrix, LIWC) have been applied to the products of student social interactions to gain a better understanding of students' (co-)construction of knowledge. These include research on the formation of student sub-communities (Yang, Wen, Kumar, Xing, & Rosé, 2014), development of self-regulation skills (Petrushyna, Kravcik, & Klamma, 2011), small-group communication (Yoo & Kim, 2013), and collaboration on computer programming projects (Velasquez et al., 2014). Further studies also investigated the link between accumulation of students' social capital in MOOCs (Dowell et al., 2015; Joksimović, Dowell et al., 2015; Joksimović, Kovanović et al., 2015), showing that position within the social network, extracted from learner interaction within various learning platforms, is associated with higher levels of cohesiveness of social media postings.

Content analytics has also been used extensively to assess the level of student engagement and instructional approaches that can contribute to its development. With this in mind, the analysis of student discussion messages – using a variety of content analytics methods – has commonly been used to assess the level of course engagement (Ramesh, Goldwasser, Huang, Daumé, & Getoor, 2013; Vega, Feng, Lehman, Graesser, & D'Mello, 2013; Wen, Yang, & Rosé, 2014b). Using probabilistic soft logic on both discussion content data and trace log data, Ramesh et al. (2013) examined student engagement in the MOOC context, focusing on the types of learners based on their level of discussion activity and course performance. Similarly, Wen, Yang, and Rosé (2014a) conducted a student sentiment analysis of MOOC online discussions, which revealed a strong association between expressed negative sentiment and the likelihood of dropping out of the course. Similar results are presented by Wen et al. (2014b) who also showed that LIWC word categories (most directly, cognitive words, first person pronouns, and positive words) could be used to measure the level of student motivation and cognitive engagement. Finally, by looking at the discrepancy between student reading time and text complexity, Vega et al. (2013) developed a content analytics system that can detect disengaged students. The general idea of using text complexity to measure engagement is that the easier the text, the shorter the reading time, unless the student is disengaged. This and similar types of analysis that combine trace data (e.g., text reading time) with the analysis of learning materials (e.g., analysis of text resource reading complexity) can be successfully used to monitor student motivation and engagement in real time, which is especially important for courses with large numbers of students, such as MOOCs.

Topic discovery in learning content

With huge amounts of web and other forms of learning data being available, one of the principal uses of content analysis is the organization and summarization of vast quantities of available information. In this regard, the most popular content analytics technique is probabilistic topic modelling, a group of methods used to identify key topics and themes in the collection of documents (e.g., discussion messages or social media posts). The most widely used topic modelling technique is latent Dirichlet allocation (LDA; Blei, 2012; Blei, Ng, & Jordan, 2003), which is often adopted in social sciences (Ramage, Rosen, Chuang, Manning, & McFarland, 2009) and digital humanities (Cohen et al., 2012). The general goal of LDA and other topic modelling techniques is to identify groups of words that are often used together, and which denote popular topics and themes in the document collection. Alongside LDA, techniques based on logic programming, text clustering, and LSA have also been used to extract main themes from student online discussions and social media postings.

Identification of main themes and topics has been extensively conducted in asynchronous online discussions. The primary goal is to raise instructors' awareness of the quality of student discourse by identifying the main themes and their magnitude in online discussions. For example, Antonelli and Sapino (2005) adopted a rule-based approach to modelling online discussions while the use of LDA has been explored by Chen (2014) and Hsiao and Awasthi (2015). In addition to topic modelling in online courses, given the large volume of discussions in massive open online courses (MOOCs), there has been particular interest in topic extraction from MOOC discussions using various approaches. Reich, Tingley, Leder-Luis, Roberts, and Stewart (2014) used structural topic models — an extension of the LDA technique that enables examining the differences in topics across different covariates — to investigate topics in MOOC online discussions and how different student (e.g., age, gender) and post characteristics (e.g., receiving an up-vote) relate to the identified topics. Likewise, Ezen-Can, Boyer, Kellogg, and Booth (2015) identified main themes in MOOC discussions through clustering "bag-of-words" representations of student online discussions.

While the discovery of topics in online discussions has been largely investigated, the analysis of main themes across different social media has received much less attention. One example is a study by Pham, Derntl, Cao, and Klamma (2012) who used SNA and word frequency analysis to investigate learning as it is occurring on popular blogging platforms and most important topics of discussion. In most of the studies, the focus of topic modelling analysis was primarily on traditional blogging platforms, while the analysis of micro-blogging platforms (e.g., Twitter) has received much less attention. In most cases, the reason for focusing on traditional blogging platforms is that most of the methods for topic modelling (e.g., LDA) are designed to work on longer text documents from which a correct topical distribution can be extracted (Zhao et al., 2011). Although several variations of LDA for short texts have been proposed (Hong & Davison, 2010; Mehrotra, Sanner, Buntine, & Xie, 2013; Ramage, Dumais, & Liebling, 2010; Yan, Guo, Lan, & Cheng, 2013), they are not currently widely used in the learning analytics field and their value is yet to be evaluated. One notable exception is the study by Chen, Chen, and Xing (2015) who — using ordinary LDA and SNA — analyzed tweets from the first four Learning Analytics and Knowledge conferences (LAK'11–LAK'14) and examined popular topics, as well as the structure and evolution of the learning analytics community over time. Similarly, a study by Joksimović, Kovanović et al. (2015) investigated the alignment between course materials and student postings in different social media (i.e., Facebook, Twitter, blogs). This study did not utilize traditional topic modelling techniques, but rather used a novel document clustering technique for topic discovery. Finally, topic modelling use has also been explored outside of social media. For example, a study by Reich et al. (2014) used LDA to examine major themes of student course evaluations, potentially providing an efficient, broad overview of course evaluation comments.

CONCLUSIONS AND FUTURE DIRECTIONS

In this chapter, we presented an overview of content analytics, a set of analytical methods and techniques for analyzing different forms of learning content in order to understand or improve learning activities. The wide range of research studies illustrates the great potential for applying content analytics techniques in addressing open problems in contemporary educational research and practice. In general, content analytics has been used for the analysis of 1) course resources, 2) student products of learning, and 3) student social interactions. Content analytics has been utilized to address a broad range of problems, such as recommendation and categorization of different learning materials (e.g., Drachsler et al., 2008), provision of feedback during student writing (e.g., Crossley et al., 2015), analysis of learning outcomes (e.g., Robinson et al., 2013), analysis of student engagement (e.g., Wen et al., 2014b), and topic discovery in online discussions (e.g., Reich et al., 2014). Given that learning analytics, as a research field, is still in its infancy, the list of problems being addressed by content analytics will likely expand in future. Likewise, as the field of content analytics matures, an important set of research practices and traditions will be established. Therefore, it is necessary to look toward future directions to provide the highest impact on educational research and practice. As such, we argue that current research in content analytics would be improved by 1) combining content analytics with other forms of analytics, and 2) developing content analytics systems based on existing educational theories. The early steps regarding the synergy between content analytics and other forms of analytics have already been observed. Several studies showed how content analytics could be successfully combined with

- **Discourse analytics** (Simsek et al., 2015, 2014, 2013),
- **Process mining** (Hever et al., 2007; Southavilay et al., 2009, 2010, 2013),
- **Social network analysis** (Drachsler et al., 2008; Joksimović, Kovanović et al., 2015; Joksimović et

al., 2014; Pham et al., 2012; Ramachandran & Foltz, 2015; Rosen et al., 2011; Teplovs et al., 2011),

- **Visual learning analytics** (Hecking & Hoppe, 2015; Lárusson & White, 2012; Pérez-Marín & Pascual-Nieto, 2010; Simsek et al., 2014; Whitelock et al., 2014, 2015), and
- **Multimodal learning analytics** (Blikstein, 2011; Worsley & Blikstein, 2011).

Likewise, it is important that additional forms of data – such as student demographics, prior knowledge, or standardized scores – are combined with content analytics, and in this regard, we also see some first steps (Crossley et al., 2015). Similar combined uses of traditional content analysis and other methods have been observed in mainstream online education research; more specifically, the use of social network analysis (De Laat, Lally, Lipponen, & Simons, 2007; Shea et al., 2010).

Finally, the development of content analytics should be based on well-established instructional theories. Many current approaches do not make use of the large body of educational research, which can limit the usefulness of the developed analytics systems and potentially even promote some detrimental learning practices (Gašević et al., 2015). Pedagogical considerations are particularly important for the provision of feedback, where the large body of previous research (Hattie & Timperley, 2007) demonstrates substantial differences in effectiveness between types of feedback provided. For example, the majority of feedback given by the current automated grading systems is summative in nature, although the most valuable feedback is on the process level, giving detailed instructions on identified weaknesses and suggestions for overcoming them. By building on existing educational knowledge, content analytics systems would not only increase in usefulness, but could also provide valuable opportunities for validation and refinement of the current understanding of learning processes.

REFERENCES

Allen, L. K., Snow, E. L., & McNamara, D. S. (2014). The long and winding road: Investigating the differential writing patterns of high and low skilled writers. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (EDM2014), 4–7 July, London, UK (pp. 304–307). International Educational Data Mining Society.

Allen, L. K., Snow, E. L., & McNamara, D. S. (2015). Are you reading my mind? Modeling students' reading comprehension skills with natural language processing techniques. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 246–254). New York: ACM. doi:10.1145/2723576.2723617

Anderson, T., & Dron, J. (2012). Learning technology through three generations of technology enhanced distance education pedagogy. *European Journal of Open, Distance and E-Learning, 2012*(II), 1–14.

Antonelli, F., & Sapino, M. L. (2005). A rule based approach to message board topics classification. In K. S. Candan & A. Celentano (Eds.), *Advances in Multimedia Information Systems* (pp. 33–48). Springer. http://link.springer.com/chapter/10.1007/11551898_6

Bateman, S., Brooks, C., McCalla, G., & Brusilovsky, P. (2007). Applying collaborative tagging to e-learning. Workshop held at the 16th International World Wide Web Conference (WWW2007), 8–12 May 2007, Banff, AB, Canada. http://www.www2007.org/workshops/paper_56.pdf

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84. doi:10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Blikstein, P. (2011). Using learning analytics to assess students' behavior in open-ended programming tasks. *Proceedings of the 1st International Conference on Learning Analytics and Knowledge* (LAK '11), 27 February–1 March 2011, Banff, AB, Canada (pp. 110–116). New York: ACM. doi:10.1145/2090116.2090132

Bosnić, I., Verbert, K., & Duval, E. (2010). Automatic keywords extraction: A basis for content recommendation. *Proceedings of the 4th International Workshop on Search and Exchange of e-le@rning Materials* (SE@M'10), 27–28 September 2010, Barcelona, Spain (pp. 51–60). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.204.5641&rep=rep1&type=pdf#page=54

Bramucci, R., & Gaston, J. (2012). Sherpa: Increasing student success with a recommendation engine. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (LAK '12), 29 April–2 May 2012, Vancouver, BC, Canada (pp. 82–83). New York: ACM. doi:10.1145/2330601.2330625

Brooks, C., Amundson, K., & Greer, J. (2009). Detecting significant events in lecture video using supervised machine learning. *Proceedings of the 2009 Conference on Artificial Intelligence in Education: Building Learning Systems That Care: From Knowledge Representation to Affective Modelling* (pp. 483–490). Amsterdam, The Netherlands: IOS Press. http://dl.acm.org/citation.cfm?id=1659450.1659523

Brooks, C., Johnston, G. S., Thompson, C., & Greer, J. (2013). Detecting and categorizing indices in lecture video using supervised machine learning. In O. R. Zaïane & S. Zilles (Eds.), *Advances in Artificial Intelligence* (pp. 241–247). Springer. http://link.springer.com/chapter/10.1007/978-3-642-38457-8_22

Buckingham Shum, S., De Laat, M. F., De Liddo, A., Ferguson, R., Kirschner, P., Ravenscroft, A., … Whitelock, D. (2013). DCLA13: 1st International Workshop on Discourse-Centric Learning Analytics. *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge* (LAK '13), 8–12 April 2013, Leuven, Belgium (pp. 282–282). New York: ACM. doi:10.1145/2460296.2460357

Buckingham Shum, S., & Ferguson, R. (2012). Social learning analytics. *Journal of Educational Technology & Society*, *15*(3), 3–26.

Buckingham Shum, S., Knight, S., McNamara, D., Allen, L., Bektik, D., & Crossley, S. (2016). Critical perspectives on writing analytics. *Proceedings of the 6th International Conference on Learning Analytics & Knowledge* (LAK '16), 25–29 April 2016, Edinburgh, UK (pp. 481–483). New York ACM. doi:10.1145/2883851.2883854

Burrows, S., Gurevych, I., & Stein, B. (2014). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, *25*(1), 60–117. doi:10.1007/s40593-014-0026-8

Calvo, R., Aditomo, A., Southavilay, V., & Yacef, K. (2012). The use of text and process mining techniques to study the impact of feedback on students' writing processes. *Proceedings of the 10th International Conference of the Learning Sciences* (ICLS '12), 2–6 July 2012, Sydney, Australia (pp. 416–423).

Cardinaels, K., Meire, M., & Duval, E. (2005). Automating metadata generation: The simple indexing interface. *Proceedings of the 14th International Conference on World Wide Web* (WWW '05), 10–14 May 2005, Chiba, Japan (pp. 548–556). ACM. http://dl.acm.org/citation.cfm?id=1060825

Charleer, S., Santos, J. L., Klerkx, J., & Duval, E. (2014). Improving teacher awareness through activity, badge and content visualizations. In Y. Cao, T. Väljataga, J. K..T. Tang, H. Leung, M. Laanpere (Eds.), *New Horizons in Web Based Learning* (pp. 143–152). Springer. http://link.springer.com/chapter/10.1007/978-3-319-13296-9_16

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2012). A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, 4(5/6), 318–331. doi:10.1504/IJTEL.2012.051815

Chen, B. (2014). Visualizing semantic space of online discourse: The Knowledge Forum case. *Proceedings of the 4th International Conference on Learning Analytics and Knowledge* (LAK '14), 24–28 March 2014, Indianapolis, IN, USA (pp. 271–272). New York: ACM. doi:10.1145/2567574.2567595

Chen, B., Chen, X., & Xing, W. (2015). "Twitter archeology" of Learning Analytics and Knowledge conferences. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 340–349). New York: ACM. doi:10.1145/2723576.2723584

Chiu, M. M., & Fujita, N. (2014a). Statistical discourse analysis: A method for modeling online discussion processes. *Journal of Learning Analytics*, *1*(3), 61–83.

Chiu, M. M., & Fujita, N. (2014b). Statistical discourse analysis of online discussions: Informal cognition, social metacognition and knowledge creation. *Proceedings of the 4th International Conference on Learning Analytics and Knowledge* (LAK '14), 24–28 March 2014, Indianapolis, IN, USA (pp. 217–225). New York: ACM. doi:10.1145/2567574.2567580

Cohen, D. J., Troyano, J. F., Hoffman, S., Wieringa, J., Meeks, E., & Weingart, S. (Eds.). (2012). Special Issue on Topic Modeling in Digital Humanities. *Journal of Digital Humanities*, 2(1).

Cook, D. A., Garside, S., Levinson, A. J., Dupras, D. M., & Montori, V. M. (2010). What do we mean by web-based learning? A systematic review of the variability of interventions. *Medical Education*, 44(8), 765–774. doi:10.1111/j.1365-2923.2010.03723.x

Corich, S., Hunt, K., & Hunt, L. (2012). Computerised content analysis for measuring critical thinking within discussion forums. *Journal of E-Learning and Knowledge Society*, 2(1), 47–60.

Crossley, S., Allen, L. K., Snow, E. L., & McNamara, D. S. (2015). Pssst... Textual features... There is more to automatic essay scoring than just you! *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 203–207). New York: ACM. doi:10.1145/2723576.2723595

Crossley, S., Roscoe, R., & McNamara, D. S. (2014). What is successful writing? An investigation into the multiple ways writers can write successful essays. *Written Communication*, 31(2), 184–214. doi:10.1177/0741088314526354

Crossley, S., Varner, L. K., Roscoe, R. D., & McNamara, D. S. (2013). Using automated indices of cohesion to evaluate an intelligent tutoring system and an automated writing evaluation system. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial Intelligence in Education* (pp. 269–278). Springer. http://link.springer.com/chapter/10.1007/978-3-642-39112-5_28

Cui, Y., & Wise, A. F. (2015). Identifying content-related threads in MOOC discussion forums. *Proceedings of the 2nd ACM Conference on Learning @ Scale* (L@S 2015), 14–18 March 2015, Vancouver, BC, Canada (pp. 299–303). New York: ACM. doi:10.1145/2724660.2728679

De Freitas, S. (2007). Post-16 e-learning content production: A synthesis of the literature. *British Journal of Educational Technology*, 38(2), 349–364. doi:10.1111/j.1467-8535.2006.00632.x

De Laat, M. F., Lally, V., Lipponen, L., & Simons, R.-J. (2007). Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for Social Network Analysis. *International Journal of Computer-Supported Collaborative Learning*, 2(1), 87–103. doi:10.1007/s11412-007-9006-4

De Wever, B., Schellens, T., Valcke, M., & Van Keer, H. (2006). Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. *Computers & Education*, 46(1), 6–28.

Di Eugenio, B., Fossati, D., Haller, S., Yu, D., & Glass, M. (2008). Be brief, and they shall learn: Generating concise language feedback for a computer tutor. *International Journal of Artificial Intelligence in Education*, 18(4), 317–345.

Donnelly, R., & Gardner, J. (2011). Content analysis of computer conferencing transcripts. *Interactive Learning Environments*, 19(4), 303–315.

Dowell, N., Skrypnyk, O., Joksimović, S., Graesser, A. C., Dawson, S., Gašević, D., ... Kovanović, V. (2015). Modeling learners' social centrality and performance through language and discourse. In O. C. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, & M. Desmarais (Eds.), *Proceedings of the 8th International Conference on Education Data Mining* (EDM2015), 26–29 June 2015, Madrid, Spain (pp. 250–258). International Educational Data Mining Society. http://www.educationaldatamining.org/EDM2015/uploads/papers/paper_211.pdf

Drachsler, H., Hummel, H. G. K., & Koper, R. (2008). Personal recommender systems for learners in lifelong learning networks: The requirements, techniques and model. *International Journal of Learning Technology*, 3(4), 404–423. doi:10.1504/IJLT.2008.019376

Dufty, D. F., Graesser, A. C., Louwerse, M., & McNamara, D. S. (2006). Assigning grade levels to textbooks: Is it just readability? In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (CogSci 2006), 26–29 July 2006, Vancouver, British Columbia, Canada (pp. 1251–1256). Austin, TX: Cognitive Science Society.

Dzikovska, M., Steinhauser, N., Farrow, E., Moore, J., & Campbell, G. (2014). BEETLE II: Deep natural language understanding and automatic feedback generation for intelligent tutoring in basic electricity and electronics. *International Journal of Artificial Intelligence in Education*, 24(3), 284–332. doi:10.1007/s40593-014-0017-9

Ellis, C. (2013). Broadening the scope and increasing the usefulness of learning analytics: The case for assessment analytics. *British Journal of Educational Technology*, 44(4), 662–664. doi:10.1111/bjet.12028

Ezen-Can, A., Boyer, K. E., Kellogg, S., & Booth, S. (2015). Unsupervised modeling for understanding MOOC discussion forums: A learning analytics approach. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 146–150). New York: ACM. doi:10.1145/2723576.2723589

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5/6), 304. doi:10.1504/IJTEL.2012.051816

Ferguson, R., & Buckingham Shum, S. (2012). Social learning analytics: Five approaches. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (LAK '12), 29 April–2 May 2012, Vancouver, BC, Canada (pp. 23–33). New York: ACM. doi:10.1145/2330601.2330616

Foltz, P. W., Laham, D., Landauer, T. K., Foltz, P. W., Laham, D., & Landauer, T. K. (1999). Automated essay scoring: Applications to educational technology. In B. Collis & R. Oliver (Eds.), *Proceedings of EdMedia: World Conference on Educational Media and Technology* 1999, 19–24 June 1999, Seattle, WA, USA (pp. 939–944). Association for the Advancement of Computing in Education (AACE). https://www.learntechlib.org/p/6607

Foltz, P. W., & Rosenstein, M. (2015). Analysis of a large-scale formative writing assessment system with automated feedback. *Proceedings of the 2nd ACM Conference on Learning @ Scale* (L@S 2015), 14–18 March 2015, Vancouver, BC, Canada (pp. 339–342). New York: ACM. doi:10.1145/2724660.2728688

Garrison, D. R., Anderson, T., & Archer, W. (2001). Critical thinking, cognitive presence, and computer conferencing in distance education. *American Journal of Distance Education*, 15(1), 7–23.

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, 59(1), 64–71. doi:10.1007/s11528-014-0822-x

Gašević, D., Mirriahi, N., & Dawson, S. (2014). Analytics of the effects of video use and instruction to support reflective learning. *Proceedings of the 4th International Conference on Learning Analytics and Knowledge* (LAK '14), 24–28 March 2014, Indianapolis, IN, USA (pp. 123–132). New York: ACM. doi:10.1145/2567574.2567590

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, 40(5), 223–234. doi:10.3102/0013189X11413260

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. doi:10.3102/003465430298487

Hecking, T., & Hoppe, H. U. (2015). A network based approach for the visualization and analysis of collaboratively edited texts. *Proceedings of the Workshop on Visual Aspects of Learning Analytics* (VISLA '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. XXX–XXX). http://ceur-ws.org/Vol-1518/paper4.pdf

Hever, R., De Groot, R., De Laat, M., Harrer, A., Hoppe, U., McLaren, B. M., & Scheuer, O. (2007). Combining structural, process-oriented and textual elements to generate awareness indicators for graphical e-discussions. In C. Chinn, G. Erkens, & S. Puntambekar (Eds.), *Proceedings of the 7th International Conference on Computer-Supported Collaborative Learning* (CSCL 2007), 16–21 July 2007, New Brunswick, NJ, USA (pp. 289–291). International Society of the Learning Sciences. http://dl.acm.org/citation.cfm?id=1599600.1599654

Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in Twitter. *Proceedings of the 1st Workshop on Social Media Analytics* (SOMA '10), 25–28 July 2010, Washington, DC, USA (pp. 80–88). New York: ACM. doi:10.1145/1964858.1964870

Hosseini, R., & Brusilovsky, P. (2014). Example-based problem solving support using concept analysis of programming content. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.), *Intelligent Tutoring Systems* (pp. 683–685). Springer. http://link.springer.com/chapter/10.1007/978-3-319-07221-0_106

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. doi:10.1145/2723576.2723613

Joksimović, S., Dowell, N., Skrypnyk, O., Kovanović, V., Gašević, D., Dawson, S., & Graesser, A. C. (2015). Exploring the accumulation of social capital in cMOOC through language and discourse. *Under Review.*

Joksimović, S., Gašević, D., Kovanović, V., Adesope, O., & Hatala, M. (2014). Psychological characteristics in cognitive presence of communities of inquiry: A linguistic analysis of online discussions. *The Internet and Higher Education*, 22, 1–10.

Joksimović, S., Kovanović, V., Jovanović, J., Zouaq, A., Gašević, D., & Hatala, M. (2015). What do cMOOC participants talk about in social media? A topic analysis of discourse in a cMOOC. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 156–165). New York: ACM.

Knight, S., & Littleton, K. (2015). Discourse centric learning analytics: Mapping the terrain. *Journal of Learning Analytics*, *2*(1), 185–209.

Kovanović, V., Joksimović, S., Gašević, D., & Hatala, M. (2014). Automated content analysis of online discussion transcripts. In K. Yacef & H. Drachsler (Eds.), *Proceedings of the Workshops at the LAK 2014 Conference* (LAK-WS 2014), 24–28 March 2014, IN, Indiana, USA. http://ceur-ws.org/Vol-1137/LA_machinelearning_submission_1.pdf

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

Krippendorff, K. H. (2003). *Content analysis: An introduction to its methodology*. Thousand Oaks, CA: Sage Publications.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284. doi:10.1080/01638539809545028

Lárusson, J. A., & White, B. (2012). Monitoring student progress through their written "point of originality". *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (LAK '12), 29 April–2 May 2012, Vancouver, BC, Canada (pp. 212–221). New York: ACM. doi:10.1145/2330601.2330653

Leeman-Munk, S. P., Wiebe, E. N., & Lester, J. C. (2014). Assessing elementary students' science competency with text analytics. *Proceedings of the 4th International Conference on Learning Analytics and Knowledge* (LAK '14), 24–28 March 2014, Indianapolis, IN, USA (pp. 143–147). New York: ACM. doi:10.1145/2567574.2567620

Manouselis, N., Drachsler, H., Vuorikari, R., Hummel, H., & Koper, R. (2011). Recommender systems in technology enhanced learning. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender systems handbook* (pp. 387–415). Springer. http://link.springer.com/chapter/10.1007/978-0-387-85820-3_12

McKlin, T. (2004). *Analyzing cognitive presence in online courses using an artificial neural network*. Georgia State University, College of Education, Atlanta, GA, United States. https://pdfs.semanticscholar.org/d6af/c0073f2efc53bb2e46a0dd39a677027b1c3d.pdf

McKlin, T., Harmon, S., Evans, W., & Jones, M. (2002, March 21). Cognitive presence in web-based learning: A content analysis of students' online discussions. *IT Forum*, 60. https://pdfs.semanticscholar.org/037b/f466c1c2290924e0ba00eec14520c091b57e.pdf

McNamara, D. S., Crossley, S., & McCarthy, P. M. (2009). Linguistic features of writing quality. *Written Communication*, *27*(1), 57–86. doi:10.1177/0741088309351547

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge, UK: Cambridge University Press.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14*(1), 1–43. doi:10.1207/s1532690xci1401_1

McNamara, D. S., Raine, R., Roscoe, R., Crossley, S., Jackson, G. T., Dai, J., … others. (2012). The Writing-Pal: Natural language algorithms to support intelligent tutoring on writing strategies. In P. M. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing: Identification, investigation and resolution* (pp. 298–311). Hershey, PA: IGI Global.

Mehrotra, R., Sanner, S., Buntine, W., & Xie, L. (2013). Improving LDA topic models for microblogs via tweet pooling and automatic labeling. *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '13), 28 July–1 August 2013, Dublin, Ireland (pp. 889–892). New York: ACM. doi:10.1145/2484028.2484166

Mintz, L., Stefanescu, D., Feng, S., D'Mello, S., & Graesser, A. (2014). Automatic assessment of student reading comprehension from short summaries. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (EDM2014), 4–7 July, London, UK. International Educational Data Mining Society. http://www.educationaldatamining.org/conferences/index.php/EDM/2014/paper/view/1372/1338

Mirriahi, N., & Dawson, S. (2013). The pairing of lecture recording data with assessment scores: A method of discovering pedagogical impact. *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge* (LAK '13), 8–12 April 2013, Leuven, Belgium (pp. 180–184). New York: ACM. doi:10.1145/2460296.2460331

Moore, M. G. (1989). Editorial: Three types of interaction. *American Journal of Distance Education*, 3(2), 1–7. doi:10.1080/08923648909526659

Muldner, K., & Conati, C. (2010). Scaffolding meta-cognitive skills for effective analogical problem solving via tailored example selection. *International Journal of Artificial Intelligence in Education*, 20(2), 99–136.

Niemann, K., Schmitz, H.-C., Kirschenmann, U., Wolpers, M., Schmidt, A., & Krones, T. (2012). Clustering by usage: Higher order co-occurrences of learning objects. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (LAK '12), 29 April–2 May 2012, Vancouver, BC, Canada (pp. 238–247). New York: ACM. doi:10.1145/2330601.2330659

Pérez-Marín, D., & Pascual-Nieto, I. (2010). Showing automatically generated students' conceptual models to students and teachers. *International Journal of Artificial Intelligence in Education*, 20(1), 47–72.

Petrushyna, Z., Kravcik, M., & Klamma, R. (2011). Learning analytics for communities of lifelong learners: A forum case. *Proceedings of the 11th IEEE International Conference on Advanced Learning Technologies* (ICALT '11), 6–8 July 2011, Athens, GA, USA (pp. 609–610). IEEE. doi:10.1109/ICALT.2011.185

Pham, M. C., Derntl, M., Cao, Y., & Klamma, R. (2012). Learning analytics for learning blogospheres. In E. Popescu, Q. Li, R. Klamma, H. Leung, & M. Specht (Eds.), *Advances in Web-Based Learning: ICWL 2012* (pp. 258–267). Springer. http://link.springer.com/chapter/10.1007/978-3-642-33642-3_28

Ramachandran, L., Cheng, J., & Foltz, P. (2015). Identifying patterns for short answer scoring using graph-based lexico-semantic text matching. *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications* (NAACL-HLT 2015), 4 June 2015, Denver, CO, USA (pp. 97–106). http://www.aclweb.org/anthology/W15-0612

Ramachandran, L., & Foltz, P. (2015). Generating reference texts for short answer scoring using graph-based summarization. *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications* (NAACL-HLT 2015), 4 June 2015, Denver, CO, USA (pp. 207–212). http://www.aclweb.org/anthology/W15-0624

Ramage, D., Dumais, S. T., & Liebling, D. J. (2010). Characterizing microblogs with topic models. In W. W. Cohen & S. Gosling (Eds.), *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media* (ICWSM '10) 23–26 May 2010, Washington, DC, USA (pp. XXX–XXX). Palo Alto, CA: AAAI Press. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1528/1846

Ramage, D., Rosen, E., Chuang, J., Manning, C. D., & McFarland, D. A. (2009). Topic modeling for the social sciences. *Workshop on Applications for Topic Models: Text and Beyond* (NIPS 2009), 11 December 2009, Whistler, BC, Canada. https://ed.stanford.edu/sites/default/files/mcfarland/tmt-nips09-20091122+21-29-34.pdf

Ramesh, A., Goldwasser, D., Huang, B., Daumé III, H., & Getoor, L. (2013). Modeling learner engagement in MOOCs using probabilistic soft logic. *NIPS Workshop on Data Driven Education* (NIPS-DDE 2013), 9 December 2013, Lake Tahoe, NV, USA. https://www.umiacs.umd.edu/~hal/docs/daume13engagementmooc.pdf

Reich, J., Tingley, D., Leder-Luis, J., Roberts, M. E., & Stewart, B. (2014). Computer-assisted reading and discovery for student generated text in massive open online courses. *Journal of Learning Analytics*, 2(1), 156–184.

Robinson, R. L., Navea, R., & Ickes, W. (2013). Predicting final course performance from students' written self-introductions: A LIWC analysis. *Journal of Language and Social Psychology*, *32*(4). doi:10.1177/0261927X13476869

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 40(6), 601–618. doi:10.1109/TSMCC.2010.2053532

Rosen, D., Miagkikh, V., & Suthers, D. (2011). Social and semantic network analysis of chat logs. *Proceedings of the 1st International Conference on Learning Analytics and Knowledge* (LAK '11), 27 February–1 March 2011, Banff, AB, Canada (pp. 134–139). New York: ACM. doi:10.1145/2090116.2090137

Roy, D., Sarkar, S., & Ghose, S. (2008). Automatic extraction of pedagogic metadata from learning content. *International Journal of Artificial Intelligence in Education*, 18(2), 97–118.

Shea, P., Hayes, S., Vickers, J., Gozza-Cohen, M., Uzuner, S., Mehta, R., ... Rangan, P. (2010). A re-examination of the community of inquiry framework: Social network and content analysis. *The Internet and Higher Education*, 13(1–2), 10–21. doi:10.1016/j.iheduc.2009.11.002

Sherin, B. (2012). Using computational methods to discover student science conceptions in interview data. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (LAK '12), 29 April–2 May 2012, Vancouver, BC, Canada (pp. 188–197). New York: ACM. doi:10.1145/2330601.2330649

Siemens, G., Gašević, D., Haythornthwaite, C., Dawson, S., Buckingham Shum, S., Ferguson, R., ... Baker, R. S. J. d. (2011, July 28). *Open learning analytics: An integrated & modularized platform.* SoLAR Concept Paper. http://www.elearnspace.org/blog/wp-content/uploads/2016/02/ProposalLearningAnalyticsModel_SoLAR.pdf

Simsek, D., Buckingham Shum, S., De Liddo, A., Ferguson, R., & Sándor, Á. (2014). Visual analytics of academic writing. *Proceedings of the 4th International Conference on Learning Analytics and Knowledge* (LAK '14), 24–28 March 2014, Indianapolis, IN, USA (pp. 265–266). New York: ACM. http://dl.acm.org/citation.cfm?id=2567577

Simsek, D., Buckingham Shum, S., Sandor, A., De Liddo, A., & Ferguson, R. (2013). XIP Dashboard: Visual analytics from automated rhetorical parsing of scientific metadiscourse. Presented at the 1st International Workshop on Discourse-Centric Learning Analytics, 8 April 2013, Leuven, Belgium. http://oro.open.ac.uk/37391/1/LAK13-DCLA-Simsek.pdf

Simsek, D., Sandor, A., Buckingham Shum, S., Ferguson, R., De Liddo, A., & Whitelock, D. (2015). Correlations between automated rhetorical analysis and tutors' grades on student essays. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 355–359). New York: ACM. http://dl.acm.org/citation.cfm?id=2723603

Snow, E. L., Allen, L. K., Jacovina, M. E., Perret, C. A., & McNamara, D. S. (2015). You've got style: Detecting writing flexibility across time. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 194–202). New York: ACM. doi:10.1145/2723576.2723592

Southavilay, V., Yacef, K., & Calvo, R. A. (2009). WriteProc: A framework for exploring collaborative writing processes. *Proceedings of the 14th Australasian Document Computing Symposium* (ADCS 2009), 4 December 2009, Sydney, NSW, Australia (pp. 129–136). New York: ACM. http://es.csiro.au/adcs2009/proceedings/poster-presentation/09-southavilay.pdf

Southavilay, V., Yacef, K., & Calvo, R. A. (2010). Analysis of collaborative writing processes using hidden Markov models and semantic heuristics. In W. Fan, W. Hsu, G. I. Webb, B. Liu, C. Zhang, D. Gunopulos, & X. Wu (Eds.), *Proceedings of the 2010 IEEE International Conference on Data Mining Workshops* (ICDMW 2010), 14 December 2010, Sydney, Australia (pp. 543–548). doi:10.1109/ICDMW.2010.118

Southavilay, V., Yacef, K., Reimann, P., & Calvo, R. A. (2013). Analysis of collaborative writing processes using revision maps and probabilistic topic models. *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge* (LAK '13), 8–12 April 2013, Leuven, Belgium (pp. 38–47). New York: ACM. doi:10.1145/2460296.2460307

Stamper, J., Barnes, T., & Croy, M. (2010). Enhancing the automatic generation of hints with expert seeding. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Intelligent tutoring systems* (pp. 31–40). Springer. http://link.springer.com/chapter/10.1007/978-3-642-13437-1_4

Strijbos, J.-W., Martens, R. L., Prins, F. J., & Jochems, W. M. G. (2006). Content analysis: What are they talking about? *Computers & Education*, *46*(1), 29–48.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29. doi:10.1177/0261927X09351676

Teplovs, C., Fujita, N., & Vatrapu, R. (2011). Generating predictive models of learner community dynamics. *Proceedings of the 1st International Conference on Learning Analytics and Knowledge* (LAK '11), 27 February–1 March 2011, Banff, AB, Canada (pp. 147–152). New York: ACM. doi:10.1145/2090116.2090139

Varner, L. K., Jackson, G. T., Snow, E. L., & McNamara, D. S. (2013). Linguistic content analysis as a tool for improving adaptive instruction. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial Intelligence in Education* (pp. 692–695). Springer Berlin Heidelberg. doi:10.1007/978-3-642-39112-5_90

Vega, B., Feng, S., Lehman, B., Graesser, A., & D'Mello, S. (2013). Reading into the text: Investigating the influence of text complexity on cognitive engagement. In S. K. D'Mello, R. A. Calvo, & A. Olney (Eds.), *Proceedings of the 6th International Conference on Educational Data Mining* (EDM2013), 6–9 July, Memphis, TN, USA (pp. 296–299). International Educational Data Mining Society/Springer.

Velasquez, N. F., Fields, D. A., Olsen, D., Martin, T., Shepherd, M. C., Strommer, A., & Kafai, Y. B. (2014). Novice programmers talking about projects: What automated text analysis reveals about online scratch users' comments. *Proceedings of the 47th Hawaii International Conference on System Sciences* (HICSS-47), 6–9 January 2014, Waikoloa, HI, USA (pp. 1635–1644). IEEE Computer Society. doi:10.1109/HICSS.2014.209

Verbert, K., Manouselis, N., Ochoa, X., Wolpers, M., Drachsler, H., Bosnic, I., & Duval, E. (2012). Context-aware recommender systems for learning: A survey and future challenges. *IEEE Transactions on Learning Technologies*, 5(4), 318–335. doi:10.1109/TLT.2012.11

Walker, A., Recker, M. M., Lawless, K., & Wiley, D. (2004). Collaborative information filtering: A review and an educational application. *International Journal of Artificial Intelligence in Education*, 14(1), 3–28.

Waters, Z. (2015). *Using structural features to improve the automated detection of cognitive presence in online learning discussions* (B.Sc. Thesis). Queensland University of Technology.

Wen, M., Yang, D., & Rosé, C. (2014a). Sentiment analysis in MOOC discussion forums: What does it tell us? In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (EDM2014), 4–7 July, London, UK. International Educational Data Mining Society. http://www.cs.cmu.edu/~mwen/papers/edm2014-camera-ready.pdf

Wen, M., Yang, D., & Rosé, C. P. (2014b). Linguistic reflections of student engagement in massive open online courses. *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media* (ICWSM '14), 1–4 June 2014, Ann Arbor, Michigan, USA (pp. 525–534). Palo Alto, CA: AAAI Press. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8057

Weragama, D., & Reye, J. (2014). Analysing student programs in the PHP intelligent tutoring system. *International Journal of Artificial Intelligence in Education*, 24(2), 162–188. doi:10.1007/s40593-014-0014-z

Whitelock, D., Field, D., Pulman, S., Richardson, J. T. E., & Van Labeke, N. (2014). Designing and testing visual representations of draft essays for higher education students. 2nd International Workshop on Discourse-Centric Learning Analytics (DCLA14), 24 March 2014, Indianapolis, IN, USA. http://oro.open.ac.uk/41845/

Whitelock, D., Twiner, A., Richardson, J. T. E., Field, D., & Pulman, S. (2015). OpenEssayist: A Supply and demand learning analytics tool for drafting academic essays. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 208–212).

New York: ACM. doi:10.1145/2723576.2723599

Wolfe, M. B. W., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). Learning from text: Matching readers and texts by latent semantic analysis. *Discourse Processes*, 25(2–3), 309–336. doi:10.1080/01638539809545030

Worsley, M., & Blikstein, P. (2011). What's an expert? Using learning analytics to identify emergent markers of expertise through automated speech, sentiment and sketch analysis. In M. Pechenizkiy, T. Calders, C. Conati, S. Ventura, C. Romero & J. Stamper (Eds.), *Proceedings of the 4th Annual Conference on Educational Data Mining* (EDM2011), 6–8 July 2011, Eindhoven, The Netherlands (pp. 235–240). International Educational Data Mining Society.

Xu, X., Murray, T., Park Woolf, B., & Smith, D. (2013). If you were me and I were you: Mining social deliberation in online communication. In S. K. D'Mello, R. A. Calvo, & A. Olney (Eds.), *Proceedings of the 6th International Conference on Educational Data Mining* (EDM2013), 6–9 July, Memphis, TN, USA (pp. 208–216). International Educational Data Mining Society/Springer.

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. *Proceedings of the 22nd International Conference on World Wide Web* (WWW '13), 13–17 May 2013, Rio de Janeiro, Brazil (pp. 1445–1456). New York: ACM.

Yang, D., Wen, M., Kumar, A., Xing, E. P., & Rosé, C. P. (2014). Towards an integration of text and graph clustering methods as a lens for studying social interaction in MOOCs. *The International Review of Research in Open and Distributed Learning*, 15(5), 214–234.

Yang, D., Wen, M., & Rosé, C. (2014). Towards identifying the resolvability of threads in MOOCs. *Proceedings of the Workshop on Modeling Large Scale Social Interaction in Massively Open Online Courses at the 2014 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2014), 25 October 2014, Doha, Qatar (pp. 21–31). http://www.aclweb.org/anthology/W/W14/W14-41.pdf#page=28

Yoo, J., & Kim, J. (2012). Predicting learner's project performance with dialogue features in online Q&A discussions. In S. A. Cerri, W. J. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Intelligent tutoring systems* (pp. 570–575). Springer. http://link.springer.com/chapter/10.1007/978-3-642-30950-2_74

Yoo, J., & Kim, J. (2013). Can online discussion participation predict group project performance? Investigating the roles of linguistic features and participation patterns. *International Journal of Artificial Intelligence in Education*, 24(1), 8–32. doi:10.1007/s40593-013-0010-8

Zaldivar, V. A. R., García, R. M. C., Burgos, D., Kloos, C. D., & Pardo, A. (2011). Automatic discovery of complementary learning resources. In C. D. Kloos, D. Gillet, R. M. C. García, F. Wild, & M. Wolpers (Eds.), *Towards ubiquitous learning* (pp. 327–340). Springer. http://link.springer.com/chapter/10.1007/978-3-642-23985-4_26

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., & Li, X. (2011). Comparing Twitter and traditional media using topic models. In P. Clough, C. Foley, C. Gurrin, G. Jones, W. Kraaij, H. Lee, & V. Murdock (Eds.), *Proceedings of the 33rd European Conference on Advances in Information Retrieval* (ECIR 2011), 18–21 April 2011, Dublin, Ireland (pp. 338–349). Springer. http://dl.acm.org/citation.cfm?id=1996889.1996934