Examining students' online interaction in a live video streaming environment using data mining and text mining

Wu He *

Department of Information Technology & Decision Sciences, College of Business and Public Administration, Old Dominion University, Norfolk, VA 23529, United States

**Abstract**

This study analyses the online questions and chat messages automatically recorded by a live video streaming (LVS) system using data mining and text mining techniques. We apply data mining and text mining techniques to analyze two different datasets and then conducted an in-depth correlation analysis for two educational courses with the most online questions and chat messages respectively. The study found the discrepancies as well as similarities in the students’ patterns and themes of participation between online questions (student–instructor interaction) and online chat messages (student–students interaction or peer interaction). The results also identify disciplinary differences in students’ online participation. A correlation is found between the number of online questions students asked and students’ final grades. The data suggests that a combination of using data mining and text mining techniques for a large amount of online learning data can yield considerable insights and reveal valuable patterns in students’ learning behaviors. Limitations with data and text mining were also revealed and discussed in the paper.

© 2012 Elsevier Ltd. All rights reserved.

**1. Introduction**

Web-based learning environments such as Blackboard and Moodle are now able to record most online learning behaviors of the students. Over the years these Web-based learning environments collect or accumulate large amounts of data and provide researchers a goldmine of unexploited data about students’ learning characteristics, behaviors, and patterns (Abdous & He, 2011; Black, Dawson, & Priem, 2008). The explosive growth in the amount of data has created a need to automatically analyze the data using novel information technology (Chiang, Lin, & Chen, 2012; Garcia, Romero, Ventura, & Gao, 2008). In recent years, there is a growing interest in applying data mining techniques to conduct the automatic analysis of learner interaction and behavioral data with web-based learning environments (Abdous & He, 2011; Garcia, Romero, Ventura, & de Castro, 2011; Hung & Crooks, 2009; Hung & Zhang, 2008; Muehlenbrock, 2005; Romero & Ventura, 2007, 2010). According to Klos- gen and Zytkow (2002), data mining is the automatic extraction of implicit and interesting patterns from large data collections. Data mining provides educational institutions the capability to explore, visualize and analyze large amounts of data in order to reveal valuable patterns in students’ learning behaviors without having to resort to traditional survey methods (Abdous & He, 2011; Talavera & Gaudioso, 2004). Turning raw data into useful information and knowledge also enables educational institutions to improve teaching and learning practices, and to facilitate the decision-making process in educational settings. Thus, educational data mining is becoming an increasingly important research area with a specific focus to exploit the abundant data generated by various educational systems for enhancing teaching, learning and decision making (Baker & Yacef, 2009; Garcia et al., 2011; Liao, Chu, & Hsiao, 2012; Romero & Ventura, 2007, 2010).

To further contribute to the understanding of educational data mining, this paper presents a study that applies data mining and text mining techniques to analyze two different data sets in a relatively new learning environment – the live video streaming (LVS) learning environment. Video streaming as a means of delivering a live course to students by computer has become an increasingly important delivery method in online learning today (Abdous & Yen, 2010; Hartsell & Yuen, 2006). The two different data sets are related to students’ social interaction in the LVS courses including online questions between students and the instructor, and online chat messages between students. More specifically, this study reveals students’ patterns of participation in two different large data sets using data and text mining techniques. In addition, to get a deeper understanding of the online participation and learning achievement, we select two LVS courses and conducted an in-depth correlation analysis. By combining both the data mining and text mining techniques, this study aims to enrich the existing body of literature, while augmenting the understanding of video-streaming (VS) students in the LVS learning environment.

*Tel.: +1 757 683 5008; fax: +1 757 683 5639.
E-mail address: whe@odu.edu
The remainder of the paper is organized as follows. Section 2 is a brief review of the literature regarding educational data mining and online social interaction. Section 3 explains the research questions explored in this paper, the context of the study, details its methodological approach (samples and procedures) as well as the key findings. Section 4 discusses the findings in depth. Section 5 discusses the instructional implications and recommendation from the results of the study. Section 6 concludes with suggestions for future research.

2. Related work

2.1. Educational data mining

According to the educational data mining community website (www.educationaldatamining.org), educational data mining (EDM) is defined to be “an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in.” Furthermore, several leading EDM experts (Baker, 2009; Baker & Yacef, 2009; Romero & Ventura, 2010) classify work in EDM into a few categories such as statistics and visualization, prediction (classification, regression, and density estimation), clustering, relationship mining, outlier detections, and text mining. EDM can be applied to assess students’ learning performance, to improve the learning process and guide students’ learning, to provide feedback and adapt learning recommendations based on students’ learning behaviors, to evaluate learning materials and courseware, to detect abnormal learning behaviors and problems, and to achieving a deeper understanding of educational phenomena (Baeppler & Murdock, 2010; Castro, Vellido, Nebot, & Mugica, 2007; Chang, 2006; Delavaria, Phon-Amnuaisuk, & Reza Beikzadeh, 2008; Faulkner, Davidson, & McPherson, 2010; Ngai, Xiu, & Chau, 2009; Romero & Ventura, 2010; Ueno, 2004; Zafra & Ventura, 2009). Text mining is focused on finding and extracting useful or interesting patterns, models, directions, trends, or rules from unstructured text documents such as such as text documents, HTML files, chat messages and emails (Abdous & He, 2011; Feldman & Dagan, 1995; Hung, 2008; Lau, Lee, & Ho, 2005; Lin, Hsieh, & Chuang, 2009; Nahm & Mooney, 2002; Romero et al., 2008). As an automated technique, text mining can be used to efficiently and systematically identify, extract, manage, integrate, and exploit knowledge for research and education (Ananiadou, 2008). Currently, there are only several studies about how to use text mining techniques to analyze learning-related data. Tane, Schmitz, and Stumme (2004) used text mining (text clustering techniques) to group e-learning resources and documents according to their topics and similarities. Hung (2012) used clustering analysis as an exploratory technique to examine e-learning literature and visualized patterns by grouping sources that share similar words, attribute values and coding rules. Some major applications of text mining include: automatic classification (clustering), information extraction (text summarization), and link analysis (Abdous & He, 2011; Hung, 2012; Perera, Kay, Koprisnka, Yacef, & Zaiane, 2009; Wetzstein et al., 2011; Xu, Wermus, & Bauman, 2011). In particular, clustering is a process of grouping objects into classes of similar objects (Jain et al., 1999; Romero et al., 2008). Clustering analysis is a well-studied technique in data mining (Lin et al., 2009) and has the advantage of uncovering unanticipated trends, correlations, or patterns from data (Chen & Liu, 2004; Duan, Xu, Guo, Lee, & Yan, 2007; Duan, Xu, Liu, & Lee, 2009; Li, Wang, & Xu, 2009; Zeng, Li, & Duan, 2012).

2.2. Online social interaction

According to Moore (1989), social interactions in class include student–instructor interaction and student–student interaction. The student–student interaction is also called peer interaction, which refers to the interaction between one student and another individual student or group of students (Moore, 1989; Zha & Ottendorfer, 2011). There are extensive researches regarding the significantly better results (26% higher marks) than those who did not, and stayed active on the site longer. Dejaeger, Goethals, Giangreco, Mola, and Baesens (2012) used different data mining techniques to identify the main drivers of student satisfaction from the data they collected in two business education institutions. The resulting models they developed provide support for the strategic decision making process.

Romero, Espejo, Zafra, Romero, and Ventura (2010) carried out several experiments and demonstrated how web usage mining can be applied in the Moodle e-learning system to predict the marks that university students will obtain in the final exam of a course. They also identified several avenues for using classification in educational settings: discovering student groups with similar characteristics, identifying learners with low motivations, proposing remedial actions, and predicting and classifying students using intelligent tutoring systems.
importance of interaction in online learning environments. Many researchers investigate the social interaction in online learning environments and in particular study the effects of interaction to learning. In general, the literature indicates that social interactions play a fundamental role in the development of cognition and make positive contributions to students’ learning (Moore, 1993; Picciano, 2002; Tu & McIsaac, 2002; Vygotsky, 1986; Zha, Kelly, Park, & Fitzgerald, 2006; Zha & Ottendorfer, 2011). For example, Tu and McIsaac (2002) indicates that an individual learner can learn more effectively through the support of others. Collaborative learning theory also places great emphasis on the extent and quality of the exchanges that occur among students in a given environment and stresses that students can broaden their knowledge base through interactions with other learners (Dillenbourg and Schneider, 1995; Dringus & Ellis, 2005; Johnson & Johnson, 1996; Macfadyen & Dawson, 2010; Smith & MacGregor, 1992; Wenger, 1998). For example, Shea, Fredericksen, Pickett, Pelz, and Swan (2001) found that the quality and quantity of interactions are important to students’ satisfaction in online courses. They conclude:

the greater the percentage of the course grade that was based on discussion, the more satisfied the students were, the more they thought they learned from the course, and the more interaction they thought they had with the instructor and with their peers. (Shea et al., 2001)

Coldwell, Craig, Paterson, and Mustard (2008) investigated the relationships between the participation and academic performance of students in an information technology course through a detailed analysis of tracking data of student participation. They found that a relationship existed between students’ participation in the online learning environment and their performance, as measured by final results in the course. The results also suggest that “the tracking data can be used as an early indicator of students who are likely to fail the course since lack of participation early in the semester is indicative of lower outcomes in the course”. They suggested that teachers and staff take remedial action proactively rather than reactively in the latter part of the semester to identify students at risks who potentially could be experiencing difficulties with their studies. In recent years, Hrastinski (2009) proposes a theory of online learning as online participation and suggests that participation and learning are inseparable and jointly constituting. This theory argues that “online learner participation is a complex process of taking part and maintaining relations with others, is supported by physical and psychological tools, is not synonymous with talking or writing, and is supported by all kinds of engaging activities”. Hrastinski (2009) further argues that “If we want to enhance online learning, we need to enhance online learner participation”. His theory indicates that students can learn from the active participation of peer students in classes (Hrastinski, 2008).

On the other hand, there is a growing body of research showing that online participation alone is not sufficient to guarantee deep and meaningful learning. The Community of Inquiry (CoI) framework developed by Garrison, Anderson, and Archer (2000) views the online learning experience as a function of three elements: social presence, teaching presence and cognitive presence. According to Garrison and Arbaugh (2007) and Swan et al. (2008),

Social presence refers to the degree to which learners feel socially and emotionally connected with others in an online environment; teaching presence is defined as the design, facilitation, and direction of cognitive and social processes for the realization of personally meaningful and educationally worthwhile learning outcomes; and cognitive presence describes the extent to which learners are able to construct and confirm meaning through sustained reflection and discourse.

Studies show that social presence is an important factor in improving instructional effectiveness and building a sense of community (Aragon, 2003; Garrison, Anderson, & Archer, 2001; Tu & McIsaac, 2002; Swan & Shih, 2005; Arbaugh, 2007; Arbaugh & Hwang, 2006). However, establishing social presence is only a pre-condition for a purposeful and worthwhile learning experience and does not guarantee that students are cognitively engaged in an educationally meaningful manner (Garrison & Cleveland-Innes, 2005). Meyer (2003) and Pawan, Paulus, Yalcin, and Chang (2003) found that quantity of interaction does not directly create cognitive development or facilitate meaningful learning and understanding, and thus does not always reflect the quality of discourse (i.e., cognitive presence). Furthermore, multiple studies show that teaching presence in the form of facilitation is crucial in the success of online learning and the instructor plays a leadership role in triggering discussion and facilitating high levels of thinking and knowledge construction (Garrison & Cleveland-Innes, 2005; Swan, 2001; Wu & Hiltz, 2004). Swan (2001) concluded that “interaction with instructors seemed to have a much larger effect on satisfaction and perceived learning than interaction with peers”. Wu and Hiltz (2004) suggest that online instructors need to structure the interaction, give more guidance and devote sufficient time to ensure that students can reach a high level of critical thinking and knowledge construction. A study by Bangert (2008) indicates strong correlations between the quality of teaching presence and student satisfaction and learning. Hung and Crooks (2009) used data mining techniques to examine and compare learning patterns between peer-moderated and teacher-moderated groups in an online learning course. The results showed that “most students in the peer-moderated condition had high participation levels and relied on student-content interaction only. Furthermore, they found that teacher presence promoted student interaction with multiple sources (content, student, and teacher). In addition, Arbaugh, Bangert, and Cleveland-Innes (2010) indicate that the CoI framework’s generalizability to online learning is mainly tested in educational disciplines and its generalizability in other disciplines is somewhat limited. A recent study by Arbaugh et al. (2010) found significant disciplinary differences (particularly regarding cognitive presence) in between applied disciplines and pure disciplines and suggest that the CoI framework may be more applicable to applied disciplines than pure disciplines. They suggest more studies are needed to examine disciplinary impacts on the CoI.

3. A research study

3.1. Research questions

While extensive researches have been done about student perceptions of their interactions and how much they have learned in Web-based courses, few studies explored the dynamics of online social interaction using data mining techniques in a LVS environment (Abdous & He, 2011). A LVS environment is different from traditional web-based learning environments such as Blackboard or Moodle. A LVS environment is designed to simulate the real classroom and provide LVS students the capabilities to watch live video lectures while being able to communicate with the instructors and other students using text messages as needed (Abdous & He, 2011). The main task for LVS students is to attend and watch the live lecture delivered on the Internet instead of online discussion. Excessive online chatting or asking too many questions could be distracting to other students who attend the class. As LVS is a relatively new technology for online learning, few prior studies have been done on the learning experience of LVS students. Ongoing evaluation and study will be necessary to achieve a deeper understanding on the use of this technology in supporting learning
as well as the effects of social interaction on students’ learning in a LVS environment.

To this end, this study applied data mining and text mining techniques to explore students’ patterns of participation and interaction in a live video streaming environment by examining the data automatically collected by the LVS environment. More specifically, the study attempts to answer the following questions:

- What are the major patterns of participation and interaction (social presence, teaching presence and cognitive presence) in LVS students’ online questions?
- What are the major patterns of participation and interaction (social presence, teaching presence and cognitive presence) in LVS students’ online chat messages?
- How is students’ participation and interaction in the LVS environment related to their learning achievement in the LVS course?

3.2. Methodology

3.2.1. Context of the study

This study was conducted in a large southeastern university which serves 24,000 students in a variety of fields. The university is also known as a national leader in technology-mediated distance learning, having served students at over 50 sites in Virginia, Arizona, and Washington state for more than 25 years. The university offers a variety of ways to deliver their courses including web-based online courses, satellite broadcasting courses, and live video-streaming courses. Using the live video-streaming (LVS) delivery mode, students participate in the class, in real time, via personal computer, over which they view a live feed of the class lecture and can interact with their instructor by sending text messages through the LVS course interface (Abdous & He, 2011). Using the same interface, LVS students are able to chat with their LVS classmates during class. At the receiving end (i.e., in the physical classroom), questions submitted by LVS students are displayed instantaneously on a monitor next to the instructor. Instructors have the option to read/answer the messages, or to save, archive, and email them for later review. This tool is intended to enable instructors to seamlessly integrate LVS students into their classroom dynamic, without distraction or overburdening during their class time (Fig. 1). The tool for sending a question to the instructor is listed in the right-top corner of Fig. 1. The chat tool for students to chat with each other is listed in the bottom of Fig. 1. A list of student participants available is shown in the right-bottom corner of Fig. 1. A big difference between the LVS software and online discussion board is that the instructor can only interact with LVS student using oral language which is broadcasted in the air. After viewing questions from LVS students, the instructor can answer the questions or ask for clarification orally in the LVS lecture.

3.2.2. Data sample

According to the university’s registrar’s office, 1144 LVS students enrolled into 138 courses in a variety of subjects (e.g., accounting, computer engineering, information technology, human services, etc.) in the Fall semester of 2009. The data sample in this study includes two different data sets of online questions and chat messages. All of the student-to-instructor questions and the two-way student-to-student chat messages were automatically recorded in text files by the LVS system. A PHP programmer wrote programs to parse these text files and extracted these questions and chat messages into a Microsoft SQL Server database and subsequently exported the data into excel spreadsheets to facilitate data processing and analysis. Data preprocessing was done to remove system testing-related messages posted by staff members. Each question or chat message in the data sets was associated with a...

![Fig. 1. LVS interface.](image-url)
course ID, a student name, a date and a time stamp. Table 1 is a description of the two different data sets.

After our data pre-processing, we identified that 298 students posted both questions and chat messages in their respective courses. We noticed that some students either posted questions or chatted with peers but failed to do both. Some students did not post anything in the LVS sessions. In addition, we obtained the final grade of these students from the university’s registrar’s office. Due to factors such as privacy and university policy, the university’s registrar’s office did not provide us with the age or gender of these students, nor could we obtain the grading scales of each course.

3.2.3. Procedures

The study includes two phases. During phase 1, we first did data pre-processing, in which raw data is transformed into a usable format, mainly by cleaning, assigning attributes, and integrating data. Subsequently, we applied various data mining and text mining techniques to examine the two different data sets in order to gain insights about students’ participation and learning behaviors. Two leading tools in textual data analysis and mining, SPSS Clementine text mining tool and NVivo 9, were used to facilitate the mining and analysis. We used the two software tools because we found each of the tools offers some advantages in certain features and functionalities. We mainly used SPSS Clementine’s linguistic methods (extracting, grouping, indexing, etc.) to explore and extract key concepts, generate categories, and help us quickly gain insights from the textual data. We mainly used NVivo 9 software to conduct various query searches and clustering analysis. The query searches are mainly used to test ideas, find interesting patterns, connections and unusual information based on the research questions. The purpose of clustering analysis in this phase is to automatically classify student posts based on the main content in their posts (questions or chat messages). The unit of analysis in our clustering analysis is based on each post (a question or chat message) from a student. A post can have more than one sentence. As Garrison, Anderson, and Archer (2010) points out, there are challenges of choosing an appropriate unit of analysis. Researchers have used grammatical units such as sentences and paragraphs, and structural units such as full postings and threads for the unit of analysis (Garrison, Cleveland-Innes, Koole, & Kappelman, 2006; Henri, 1991; Zha & Ottendorfer, 2011). Different ways have different strengths and weaknesses (e.g., reliability issues). As this study is proposing to use text mining approach to automatically analyze the texts without resorting to the traditional manual coding, we found that a message-level unit was the most appropriate for our goal. Garrison et al. (2001) also suggest that the message as unit is “attractive because the length and content of the message is decided upon by its author, rather than by coders”. Thus, we associated each post

<table>
<thead>
<tr>
<th>Data sets</th>
<th>Total numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online questions (student-to-instructor)</td>
<td>3678 questions (from 84 courses posted by 339 students)</td>
</tr>
<tr>
<td>Online chat messages (student-to-student)</td>
<td>29,828 chat messages (from 115 courses posted by 825 students)</td>
</tr>
</tbody>
</table>

Fig. 2. Data/text mining process for a LVS system (adapted from Abdous and He (2011).
Major Themes in students' online questions.

Table 2

<table>
<thead>
<tr>
<th>#</th>
<th>Main theme</th>
<th>Frequency ($)</th>
<th>Text content</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Check-in/check out</td>
<td>41</td>
<td>Check-in and checking out messages</td>
</tr>
<tr>
<td>2</td>
<td>Social and affective</td>
<td>22</td>
<td>Greetings, giving thanks, and emotional responses</td>
</tr>
<tr>
<td></td>
<td>statements</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Learning/comprehension</td>
<td>15</td>
<td>Questions regarding course materials and assignments (e.g., asking instructor to go over a concept or assignment)</td>
</tr>
<tr>
<td>4</td>
<td>LVS technical issues</td>
<td>10</td>
<td>Telling instructors that they had or are having technical problems (video, audio, text size on screen, networking, LVS software, browser compatibility issues, etc.); Reporting that technical problems are solved.</td>
</tr>
<tr>
<td>5</td>
<td>Course logistics</td>
<td>9</td>
<td>Submission deadline, exam schedule, lab schedule, Exam format, assignment requirements, grading, office hours</td>
</tr>
<tr>
<td>6</td>
<td>Others</td>
<td>3</td>
<td>Messages that are difficult to categorize</td>
</tr>
</tbody>
</table>

3.3. Findings

The main findings of the study are summarized as below:

- What are the major patterns of participation and interaction (social presence, teaching presence and cognitive presence) in LVS students' online questions?

The online question data set includes 3678 questions posted by 339 LVS students in 84 courses. Descriptive statistical analysis was done to get a basic understanding of these questions. The average number of questions posted by a student is 10.8 (standard deviation = 13.8; range = from 1 to 118). The average number of questions in a LVS course is 44 (standard deviation = 89; range = from 2 to 548). The average number of words contained in a question is 21 words (range from 1 word to 105 words). In an extreme case, a student's question contains eight short sentences.

SPSS Clementine extracted many concepts from the data set of online questions and can sort the extracted concepts by their frequency. The most outstanding concept identified is checking-in. Many VS students sent a checking-in message to the instructor to indicate that they were participating and watching the live video at the beginning of each LVS session. Some students explained in their checking-in messages why they were late for checking-in (e.g., trouble with the Internet or computers). Other top key concepts include problems (videos, sound, login, etc.), schedule (test, exam, lab, assignment or class), homework/assignment questions, etc. Next, we applied several concept-grouping techniques (concept derivation, concept inclusion and semantic networks) to generate categories by using Clementine. In the meantime, we used NVivo 9 to run query searches and a clustering analysis for all the questions, which generated a multi-level clusters based on the word similarities. After reviewing the generated categories and clusters, we made some adjustments by deleting, merging, combining and/or refining. Finally, we identified a few main categories/themes from all the online questions (see Table 2). The main categories/themes, their frequencies and simple descriptions are listed below.

From the perspective of the Community of Inquiry (CoI), text mining was able to quickly identify evidences related to social presence, teaching presence and cognitive presence from these online questions but was not able to quickly tell the percentage of their presence. For example, many key concepts extracted from the questions are related to social presence. After examining the text associated with these extracted key concepts, we found many examples of social presence include check-in messages, greeting the instructor or giving thanks to the instructor, expressing emotions and feelings (happiness or confusion), and telling the instructor about the technical issues or personal issues (sickness, traffic delay, work delay, etc.). There were some concepts related to cognitive presence. For example, some subject matter related concepts such as workforce, network configuration, and feminist movement appeared and they were related to cognitive presence. Students either asked quite a few course content related questions or answered the instructor's questions. But a deep examination of these evidences shows that there are only a few questions that belong to high level of critical thinking and knowledge construction. The majority of cognitive-related questions are low level cognitive interaction. As instructors only interact with LVS students using oral language and do not post textual data at all through the LVS software, teaching presence was only indirectly reflected through students' post. We found many posts that responded to instructor's questions or assigned tasks and activities. For example, some LVS students are not sure about the in-class exercises given by the instructor and asked the instructor to provide more information about the exercises. Although a number of concepts were extracted and identified with links to the evidences in the text collection, a limitation with current text mining techniques is that it could not easily show us the percentage of these different presences unless researchers took a lot of time to manually code the messages and count them. We hope that future research on text mining can make breakthrough on this aspect and allows researchers to easily find the percentage based on their query for selected concepts.

We also examined the disciplinary effects in online LVS participation using descriptive statistics. The results identified discipline-based differences in terms of the average questions asked by LVS students from different colleges. As Fig. 3 shows, LVS students in Health Sciences, Education and Arts & Letters clearly asked significantly more questions than students in other colleges...
(including Business & Public Administration, Engineering & Technology, and Sciences). In comparison, students in Engineering and Science asked questions infrequently. This finding indicates a need to conduct additional research for developing discipline-based models to explain student behaviors in LVS courses. Some possible explanations for this finding are presented in the discussion section.

- What are the major patterns of participation and interaction (social presence, teaching presence and cognitive presence) in LVS students’ online chat messages?

The online chat message data set includes 29,828 chat messages posted by 825 students in 115 courses. We conducted descriptive statistical analysis for these chat messages as well. The average number of chat messages posted by a student is 36 (standard deviation = 53.9; range = from 1 to 528). The average number of chat messages in a LVS course is 259 (standard deviation = 335; range = from 1 to 1664). The average number of words contained in a chat message is appropriately 12 words (range from 1 word to 60 words).

We followed the same approach to analyze the chat messages using SPSS Clementine. As a result, five main categories/themes were identified from these online chat messages. The main categories/themes are listed in Table 3. Due to a technical issue with clustering analysis, we were not able to get a percentage for these themes.

An interesting finding is that many students used the chat tool to troubleshoot the LVS technical issues and help each other. For example, many students chose to ask their online peers first when they had LVS related issues such as audio or video problems instead of contacting the instructor or technical support. As a result, LVS students often help each other and provide comments or suggestions for troubleshooting.

We tried to use NVivo 9 cluster these chat messages. Unfortunately, due to the large size of the data set, NVivo was not able to generate clusters efficiently and caused our computers (two computers with fast CPU and large memory) to freeze. Thus, we had to give up this attempt. As a result, we were not able to get the percentage for the above themes. This issue reveals a practical limitation with existing text analysis/mining tools. A labor-intensive coding process will have to be conducted. Thus, at present we can only provide some examples and observations related to these themes.

Similarly, the current text mining techniques could not give the percentage of their presence in chat messages, a labor-intensive coding process will have to be conducted. Thus, at present we can only provide some examples and observations related to these themes here. We hope that text mining tools can overcome this limitation in the future. In summary, we observed abundant social exchange messages between students such as greetings, introducing to each other, helping each other with LVS issues, talking about school, and personal interests or news. There were also a lot of messages involving critical thinking and knowledge construction. For example, a student helps a classmate to debug an error in his program; several students discussed how to address a problem in doing the group project and showed their critical thinking and reasoning skills. We also found a few evidences that indirectly reflected teaching presence. For example, an LVS student shared his opinion on a specific task the instructor just assigned in the lecture and asked other LVS students to provide more input for this task in order to generate a group summary.

We also examined the disciplinary effects in online LVS participation in terms of the average chat messages posted by LVS students from different colleges. As Fig. 4 shows, LVS students in Education and Health Sciences clearly chatted more often than students in other colleges. In contrast, students in Engineering and Science chatted infrequently. This specific finding is quite consistent with the disciplinary effects we found for online questions. Further research is definitely needed to examine why LVS students in Engineering and Science are not active in LVS sessions. Some possible explanations for this finding are presented in the discussion section.

- How is students’ participation and interaction in the LVS environment related to their learning achievement in the LVS course?

After conducting a mining of the data set as a whole, we decided to narrow down the analysis and selected two courses for an in-depth examination in order to get more specific understanding how students’ participation and interaction may affect student learning outcomes. One course with the most questions and one course with the most chat messages were selected for the in-depth examination. Both courses happen to be educational courses of-
fered by the College of Education. Both courses are focused on conceptual understanding of different educational issues and educational theories. Course A is a senior-level course about instructional strategies for education. Course B is a senior-level course introducing general and special education. Table 4 describes some basic data of the two courses. We were able to acquire students’ final grades for the selected courses from the university’s registrar’s office too. As Table 4 shows, each student in course A posted nearly 46 questions and 59 chat messages on average during the semester; each student in course B posted about 11 questions and 52 chat messages on average during the semester.

First, we calculated a correlation between the number of questions, number of chat messages and students’ final grade for each course using SPSS (Norusis, 2008). The results were summarized in Table 5. No correlation was found between the number of questions and chat messages in the two courses. However, we found a positive correlation between the number of questions students asked and their final grade for both courses as Table 5 shows. A negative correlation was found between the number of chat messages and final grade for course B.

As a strong positive correlation and a moderate positive correlation between the number of questions and final grades were found, we were interested in learning more about the underlying causes. Next, we manually classified the online questions of the course using the six main themes identified in Table 2. When a question includes multiple aspects, we used the “code up” approach (Garrison et al., 2001) and conducted classification based on the most important meaning of the question (Henri, 1991). In this way, each question is classified into one category. The results were also compared, reviewed and discussed to ensure the accuracy.
racy and reliability. After the classification is complete, we run correlation analysis again using SPSS to check the correlations between different themes and the final grade. As a result of the correlation analysis, we found:

- There is a moderate positive correlation \((r = 0.39)\) between the number of check in/check out messages and the final grade in course A. There is a weak positive correlation \((r = 0.24)\) between the number of course content learning/comprehension and the final grade in course A. We did not find correlations between other themes and the final grade in course A.
- There is a strong positive correlation \((r = 0.43)\) between the number of check in/check out messages and the final grade in course B. We did not find correlations between other themes and student's final grade in course B.

4. Discussion

The findings reveal discrepancies as well as similarities in students' patterns of participation between online questions (student–instructor interaction) and online chat messages (student–students interaction or peer interaction). Students were much more active in interacting with their peers than with their instructors by comparing the numbers of questions and chat messages they created during LVS sessions. For example, there were a lot more chat messages than questions talking about LVS-related technical problems (videos, sound, etc.). Instead of interrupting the instructor's teaching with technical issues, may LVS students choose to ask their peers for testing and help first. The findings also show that the most outstanding theme in online questions is about course check-in and check-out. Many students regularly sent “check-in” messages to the instructor to indicate that they were watching the live class lecture on the Internet. A possible explanation is that instructors require LVS students to check in as a part of their course participation grade. In addition, we also found that a few students in some courses did not check in or ask instructors questions at all during the LVS sessions. The underlying reason is unknown and needs further investigation. However, this finding corroborates the opinions of Kreijns, Kirschner & Jochems (2003) that we cannot take for granted that interaction automatically takes place just because an environment makes it technologically possible.

In addition, we found that the most outstanding theme in the chat message data set is related to social and affective statements. It seems that students were more willing to share information and feelings with their peers and discussed their concerns with peers. For example, some students discussed what courses they were taking and what courses they will be taking next semester. They also shared personal hobbies such favorite movies or games with the peers. Instead of asking the instructor the submission deadlines or requirements or repeat something, they were more likely to check with their peers first before asking the instructor. For example, a student asked the peers that “Morning, getting a late start, anything special on the agenda?” They also asked the peers their opinions about particular assignments or exams. For example, questions like “how did you all feel about the test” appeared many times in the chat message data set. Students also help each other for some of the assignment questions. In a programming course, a student posted a programming error and got responses and specific suggestions from other students. When there is a technical problem with the LVS system, students also tend to check with other students and see if they had the same problem. The chat messages also show that students were eager to help each other and reported solved technical issues. Below is an example scenario:

- Is your video streaming working fine? I have no sound yet.
- Try F5 if no sound.
- Yeah! I had to hit it like 10 times to get it to work.

In addition, we found that students are more likely to use the chat messages to communicate positive emotions, negative emotions, and expressions of social support. For example, a few students told other students how frustrated they felt about the assignments or exams. Other expressed emotional support and suggested ways to deal with the issues.

The findings also reveal a disciplinary difference in online LVS participation by LVS students from different colleges. The results show that students in Engineering and Science asked significantly less questions than students in Education and Health Science. Similarly, students in Engineering and Science also chatted much less frequently than students in Education and Health Science. This interesting finding suggests a need for further investigation into the effect of disciplinary differences on LVS students' online interaction. Previous studies (Arbaugh et al., 2010; Neumann, Parry, & Becher, 2002) indicate that in “hard” disciplines, teaching activities are more focused and instructive, with the primary emphasis being on the teacher informing the student. In contrast, teaching and learning activities in “soft” disciplines tend to be more constructive, reiterative and reflective. In “soft” disciplines, essays, group projects, reflection and discussion are more commonly used for knowledge building than “hard” disciplines. The disciplinary difference between hard and soft disciplines may provide possible explanations for the differences in LVS students' online participation and interaction from different colleges. The results of this study also reveal a need to develop discipline-specific instructional approaches for better engaging students.

In addition, our analysis about two specific courses found positive correlations between the number of questions students sent to the instructor and their final grades. Positive correlations between the number of check in/check out messages and the final grade were also found in both courses. It seems that students who often ask questions and regularly attend the class online are more likely to achieve better grades. The findings are reasonable and support some previous findings in online studies (Coldwell et al., 2008; Shea et al., 2001). In addition, the findings in course A reveal that those who often ask most questions tend to not only check in more frequently than those who asked few questions, but also ask many more course content related questions than those who asked few questions. It seems that those more active students paid more attention to the course content itself than those less active students in course A. This might explain why students who asked more questions tend to have better performance in the course. However, it is also possible that students' academic competence influenced their overall understanding about the course content and consequently affected their interaction with the instructor. Further studies are needed to identify the causal relationship between students' participation and cognitive achievement in the student-to-instructor interaction and their learning outcome in the course (Zha & Ottendorfer, 2011). In addition, as the findings were found in educational courses, such correlations may not exist in other subjects because of disciplinary differences and other factors such as different instructional approaches and grading criteria. Thus, the findings should not be generalized and more future research is needed to test the effect of students' interaction with the instructor on their final grades in different course contexts.

On the other hand, we did not find a positive correlation between the number of chat messages posted by students and their final grades. We even found a negative correlation between the number of chat messages posted by students and their final grades for course B. As a LVS environment is different from traditional
online discussion board, LVS students need to focus on watching the live video lectures and interact with the instructors as needed. Excessive chatting with the peers may distract student attentions and affect their learning effects in the live class session. Some students probably realized this and did not chat with their peers unless it was definitely needed. But some students did not realize this and chatted a lot with their peers online. Thus, the instructor better remind their LVS students not to chat excessively about stuff that is not related to the course. In addition, the literature shows that student’s interaction with other students could be problematic. For example, Sproull and Kiesler (1991) identified an issue caused by discussions that continue based on misinformation (in some cases an instructor cannot immediately correct or clarify a misleading comment). In addition, the amount of student interaction and the number of comments can easily lead to information overload issues (Jones, Ravid, & Rafaeli, 2004; Vonderwell & Zachariah, 2005). Furthermore, Garrison and Cleveland-Innes (2005) suggest that surface exchange of information cannot guarantee deep and meaningful learning. These factors may explain why a lot of chatting with other students in LVS environments either does not help their final grades or could even cause negative effects.

A limitation with the study is that it only examined online questions and chat messages collected by the LVS system. Future research needs to look at the entire courses (Shea, Hayes, & Vickers, 2010) and not just at the LVS questions or chat messages. A holistic view of student demographic and institutional variables, as opposed to the single variable, must also be examined in determining the overall online learning experience including students’ perception and satisfaction (Garrison & Cleveland-Innes, 2005; Herbert, 2008; Shea & Bidjerano, 2009). However, the evidences and findings revealed in this study do show the benefits and value of data and text mining in researching learners’ behavior in online learning environments. While data mining and text mining have many potential benefits and values, there are still some technological limitations to their capabilities.

- One limitation is that a post (a unit of the analysis) may contain multiple cognitive aspects or contradictory categorization cues. However, the clustering analysis used in text mining only assign a post to one cluster based on one of its content characteristics (word similarities as defined in the dictionary) and will not assign a post to multiple clusters based on its multiple meanings/themes. This will result in missing data or value for certain clusters.
- A second limitation is that while data and text mining can reveal hidden insights, patterns, and relationships, they cannot tell the user their value or significance of the insights, patterns, and relationships (Seifert, 2004). A lot of manual interpretation and analysis is still required to support decision making.
- A third limitation is that data and text mining cannot identify a causal relationship between behaviors and variables (Seifert, 2004). With continuous technological advances in the data and text mining area, we hope that these technical limitations would be mitigated or resolved in the near future.

5. Implications and recommendations

The results reveal that many students encounter technical issues and problems from time to time during the semester. Thus, there is a need for the LVS practitioners to constantly monitor student participation and patterns of participation in order to identify, solve and prevent technical issues and problems (Abdous & He, 2011). On the other hand, these technical issues and problems provide an opportunity for further research and development of the LVS system. Our findings can be used to inform the maintenance and future development of the LVS software. Firstly, LVS software developer or technical support personnel can use text mining to identify potential technical issues students had with the LVS system and thus to take a proactive approach to do troubleshooting and system maintenance. Secondly, the current version of the LVS software only shows a list of logged-in students but does not show the logged-in students’ presence status (e.g., busy, away, available, appear offline). There is a need for the LVS software developers to implement features that can display the presence status of online students since social presence is vital to online interaction (Tu & McIsaac, 2002). Thirdly, LVS software developers need to implement features to give instructors graphical representations about student participation. Currently, the LVS system only provides text information (questions and chat messages) to instructors. A graphical LVS student tracking and monitoring tool is needed to extract the recorded data by the system and to generate graphical representations that can be easily explored by instructors to examine various aspects about LVS students. For example, graphs reporting all students’ login information (login time and duration) to the LVS course and reporting their participation information (questions and chat messages over time) would be allowing instructors to gain a better understanding of their students’ online behavior and become aware of what is happening in LVS sessions (Romero et al., 2008).

In addition to the technical issues and problems, a common challenge for online learning is to encourage learner participation (Hrastinski, 2008). This study introduces the techniques that LVS practitioners and decision makers can use to identify students at risks with low or no participation. Instructors can also use these techniques to gain insights from students’ questions and chat messages in order to adjust instructional methods and scaffold learning accordingly. For example, the findings indicate that many instructors did not take sufficient efforts to engage LVS students in their courses and students in some disciplines only asked a few questions during the entire semester. There is a need for instructors to adopt more strategies and ways to encourage participation and engage students to ask questions related to course content learning and comprehension. Some recommendations to improve instructional effectiveness and to improve LVS student learning experience are listed below:

- A training session for new instructors who teach LVS courses should be required. In particularly, institutions should provide training and support for instructors who teach LVS courses around teaching presence (Abdous & He, 2011; Shea et al., 2010). A training session can teach instructors useful discipline-specific facilitation techniques to engage LVS students for deep learning and knowledge construction in LVS courses.
- Instructors need to require LVS students to check-in and check-out for each LVS class session to ensure that they are watching the live video-streamed lecture. In addition, instructors should encourage students to make a self-introduction at the beginning of each semester so that students can know each other. An online page with student profile information should be included to help build familiarity among students (Aragon, 2003).
- Instructors should provide LVS students with tips on effective participation and interaction during LVS sessions (writing messages, expressing emotions and feelings, communication style and strategy, timing of questions, length of messages, timely response to messages, etc.). Instructors are recommended to make clear to their LVS students to what extent and in what capacity they will participate in course discussions (Shea et al., 2010). Instructors also need to remind students that excessive chatting online could distract other LVS students and affect other students’ learning from the live video-streamed lecture.
Instructors should provide LVS students with collaborative learning activities and teaching tasks from time to time to motivate online interaction, encourage deep learning, and help build familiarity, trust and other relationships among LVS students. Tu and McIsaac (2002) suggest that several task types including planning, creativity, intellectual, decision making, cognitive conflict, and social tasks can influence online interaction and student learning.

Instructors should assign a certain percentage of course grades based on LVS students’ participation and interaction in LVS sessions. Instructors need to let students know that the actual points will be determined by both quantity and quality of their participation which are characterized by the questions and chat messages.

Instructors should communicate privately with students who are ineffective in their postings or who fail to participate (Shea et al., 2010). Instructors can use the existing summary report of students questions and chat messages provided by the LVS system to identify students at risk and then contact them to provide specific guidance.

6. Conclusions and future research

This paper discusses methods to exploit large amounts of untapped student data including online questions and chat messages automatically collected by a LVS system. The data and text mining results revealed some interesting patterns and themes in student’s interaction with the instructor and other students. Furthermore, this study reveals a disciplinary difference in online LVS participation by LVS students from different colleges. The results show that students in Engineering and Science asked significantly less questions than students in Education and Health Science. Similarly, students in Engineering and Science also chatted much less frequently than students in Education and Health Science. A further examination about two educational courses found positive correlations between the number of questions students sent to the instructor and their final grades. These findings suggest interesting opportunities for future researchers to conduct disciplinary studies in LVS learning environments.

The study clearly shows the value of educational data mining and text mining as an alternative analytical approach in gaining insights from large amounts of untapped textual data. Compared to traditional time- and labor-intensive human content analysis of textual information, data and text mining can effectively reduce the use of time and manual labor in identifying insights and patterns from large text collection and hold great potential in online learning analysis. The study also makes a contribution to the understanding of students’ online patterns of participation in a relatively new LVS environment. Instead of using traditional survey or interview approaches, the study uses a novel approach of combining data mining and text mining techniques to explore the LVS students’ online patterns of participation. This study added new evidences to the knowledge base of online learning and provides a basis for conducting a holistic study to understand LVS students’ learning behavior and perceptions. A main limitation of the current study is that we relied on a single method to examine the data automatically collected by the LVS system to gain insights. Additional methods such as surveys and interviews are also needed to probe more deeply about students’ perceptions, thoughts, experiences and cognitive processes. We plan to develop a questionnaire to survey students for some issues we found through the data and text mining process. For example, we want to know why some students participated in the online chat but did not ask instructor questions. We also want to learn about students’ perceptions about the effect of peer interaction on their learning in the LVS environment. A longitudinal study involving multiple semesters is also needed to further verify and assess LVS students’ patterns of participation and learning behaviors over a long period of time. Further investigation of students’ engagement and the dynamics of their interaction in these new technology-based learning environments will contribute to the development and improvement of technology-oriented learning environments and online teaching & learning practices.

Acknowledgements

The author would like to thank the anonymous referees for their constructive comments which led to a significant improvement of this paper. The author would also like to thank M’Hammed Abdous, Chering-Jhy Yen, Xin Tian, Yao Li and Shawn Altman for their assistance with the data collection and analysis.

References


