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

ARTICLE INFO

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.

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, 2011; Duan, Street, & Xu, 2011; Li, 2011; Shi et al., 2007; Xu, Liang, & Gao, 2008). In recent year, 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 Klosgen 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 videostreaming (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 (Baepler & Murdoch, 2010; Castro, Vellido, Nebot, & Mugica, 2007; Chang, 2006; Delavaria, Phon-Amnuaisuka, & Reza Beikzadehb, 2008; Faulkner, Davidson, & McPherson, 2010; Ngai, Xiu, & Chau, 2009; Romero & Ventura, 2010; Romero, Ventura, & Garcia, 2008).

Quite a few EDM studies have been found in the most recent literature from 2010 to 2012. For example, Pal (2012) used machine learning algorithm, a data mining technique, to analyze and extract information from existing student data to establish a predictive model. The predictive model is then used to find students which are likely to drop out their first year of engineering. Similarly, Jovanovica, Vukicevica, Milovanovica, and Minovica (2012) applied classification models for predicting students' performance, and cluster models for grouping students based on their cognitive styles in e-learning environment. They indicate that the classification models helped teachers, students and business people, for early engaging with students who are likely to become excellent on a selected topic. Furthermore, they indicate that clustering students based on cognitive styles and their overall performance enable better adaption of the learning materials with respect to their learning styles. Along the same lines, Parack, Zahid, and Merchant (2012) used multiple data mining algorithms for student profiling and grouping. They found that data mining can be very useful in discovering valuable information which can be used for profiling students based on their academic record such as exam scores, term work grades, attendance and practical exams.

Falakmasir and Jafar (2010) applied data mining methods to the web usage records of students' activities in the Moodle virtual learning environment. As a result, they were able to identify and rank the students activities based on their impact on the performance of students in final exams. Their findings suggest that students' participation in virtual classrooms had the greatest impact on their final grades. Dominguez, Yacef, and Curran (2010) created a system to generate personalized feedback and hints by mining the student data collected by an online learning system. They found that students who used the hinting system achieved

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.

Gorissen, Bruggen, and Jochems (2012) analyzed the interactions of students with the recorded lectures using educational data mining techniques. The data logged by the lecture capture system (LCS) was used and combined with collected survey data. They found discrepancies as well as similarities between students' verbal reports and actual usage as logged by the recorded lecture servers. The data suggests that students who do this have a significantly higher chance of passing the exams. They concluded that given the discrepancies between verbal reports and actual usage, research should no longer rely on verbal reports alone.

An emerging trend in EDM is the use of text mining which is an extension of data mining to text data (Ananiadou, 2008; Barahate & Shelake, 2012; Liu, Cao, & He, 2011; 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, Koprinska, 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