2012

Using Data Mining for Predicting Relationships between Online Question Theme and Final Grade

M'hammed Abdous
Old Dominion University

Wu He
Old Dominion University

Cherng-Jyh Yen
Old Dominion University

Repository Citation

Abdous, M'hammed; He, Wu; and Yen, Cherng-Jyh, "Using Data Mining for Predicting Relationships between Online Question Theme and Final Grade" (2012). *Information Technology & Decision Sciences Faculty Publications*. 21.
https://digitalcommons.odu.edu/itds_facpubs/21

Original Publication Citation

Abdous, M., He, W., & Yen, C. J. (2012). Using data mining for predicting relationships between online question theme and final grade. *Educational Technology & Society, 15*(3), 77-88.

Abdous, M., He, W., & Yen, C.-J. (2012). Using Data Mining for Predicting Relationships between Online Question Theme and Final Grade. *Educational Technology & Society*, *15* (3), 77–88.

Using Data Mining for Predicting Relationships between Online Question Theme and Final Grade

M'hammed Abdous[1], Wu He[2*] and Cherng-Jyh Yen

[1]Center for Learning and Teaching // [2]Information Technology Department of Information Technology & Decision Sciences // [3]Educational Research and Statistics Department of Educational Foundations and Leadership, Old Dominion University, Norfolk, Virginia, 23529, USA // mabdous@odu.edu // whe@odu.edu // cyen@odu.edu

[*]Corresponding Author

ABSTRACT

As higher education diversifies its delivery modes, our ability to use the predictive and analytical power of educational data mining (EDM) to understand students' learning experiences is a critical step forward. The adoption of EDM by higher education as an analytical and decision making tool is offering new opportunities to exploit the untapped data generated by various student information systems (SIS) and learning management systems (LMS). This paper describes a hybrid approach which uses EDM and regression analysis to analyse live video streaming (LVS) students' online learning behaviours and their performance in their courses. Students' participation and login frequency, as well as the number of chat messages and questions that they submit to their instructors, were analysed, along with students' final grades. Results of the study show a considerable variability in students' questions and chat messages. Unlike previous studies, this study suggests no correlation between students' number of questions / chat messages / login times and students' success. However, our case study reveals that combining EDM with traditional statistical analysis provides a strong and coherent analytical framework capable of enabling a deeper and richer understanding of students' learning behaviours and experiences.

Keywords

Educational data mining, Data mining, Live video streaming, Clustering analysis

Introduction

According to a recent survey conducted by Campus Computing (campuscomputing.net) and WCET (wcet.info), almost 88% of the surveyed institutions reported having used an LMS (Learning Management System) as a medium for course delivery for both on-campus and online offerings. In addition to various student information management systems (SISs), LMSs are providing the educational community with a goldmine of unexploited data about students' learning characteristics, behaviours, and patterns. The turning of such raw data into useful information and knowledge will enable institutes of higher education (HEIs) to rethink and improve students' learning experiences by using the data to streamline their teaching and learning processes, to extract and analyse students' learning and navigation patterns and behaviours, to analyse threaded discussion and interaction logs, and to provide feedback to students and to faculty about the unfolding of their students' learning experiences (Hung & Crooks, 2009; Garcia, Romero, Ventura, & de Castro, 2011). To this end, data mining has emerged as a powerful analytical and exploratory tool supported by faster multi-core 64 CPUs with larger memories, and by powerful database reporting tools. Originating in corporate business practices, data mining is multidisciplinary by nature and springs from several different disciplines including computer science, artificial intelligence, statistics, and biometrics. Using various approaches (such as classification, clustering, association rules, and visualization), data mining has been gaining momentum in higher education, which is now using a variety of applications, most notably in enrolment, learning patterns, personalization, and threaded discussion analysis. By discovering hidden relationships, patterns, and interdependencies, and by correlating raw/unstructured institutional data, data mining is beginning to facilitate the decision-making process in higher educational institutions.

This interest in data mining is timely and critical, particularly as universities are diversifying their delivery modes to include more online and mobile learning environments. EDM has the potential to help HEIs understand the dynamics and patterns of a variety of learning environments and to provide insightful data for rethinking and improving students' learning experiences.

This paper is focused on understanding live video streaming (LVS) students' learning behaviours, their interactions, and their learning outcomes. More specifically, this study explores how the interaction of students with each other and with their instructors predicts their learning outcomes (as measured by their final grades). By investigating these interrelated dimensions, this study aims to enrich the existing body of literature, while augmenting the understanding of effective learning strategies across a variety of new delivery modes.

This paper is divided into four sections. It begins by reviewing the literature dealing with the use of data mining in administrative and academic environments, followed by a short discussion of the way in which data mining is used to understand various dimensions of learning. The second section explains the purpose and the research questions explored in this paper. The third section describes the background of the study and details its methodological approach (sampling, data collection, and analysis). The paper concludes by highlighting key findings, by discussing the study's limitations, and by proposing several recommendations for distance education administrators and practitioners.

Data mining applications in administrative and academic environments

At the intersection of several disciplines including computer science, statistics, psychometrics (Garcia et al., 2011), data mining has thrived in business practices as a knowledge discovery tool intended to transform raw data into high-level knowledge for decision support (Hen & Lee, 2008). To this end, a wide range of tools that can be used for collecting, storing, analysing, and visualizing data, such as the SPSS Modeler (formerly Clementine) and the SAS Enterprise Miner, have been developed in the business world. These tools use sophisticated computing paradigms including decision tree construction, rule induction, clustering, logic programming, and statistical algorithms.

Although data mining has been widely used in business environments to predict future trends and consumer behaviours (Harding, Shahbaz, Srinivas, & Kusiak, 2006; Ngai, Xiu, & Chau, 2009), the data mining method has been dramatically under-used in education research in general (Faulkner, Davidson, & McPherson, 2010). Only recently have higher education institutions started to exploit the potential of this powerful analytical tool (Black, Dawson, & Priem, 2008).

However, according to Romero and Ventura (2010), educational data mining (EDM) has emerged as a new field of research capable of exploiting the abundant data generated by various systems for use in decision making. The enthusiastic adoption of data mining tools by higher education has the potential to improve some aspects of the quality of education, while it lays the foundation for a more effective understanding of the learning process (Baker & Yacef, 2009). EDM, when integrated into an iterative cycle (Romero, Ventura, & Garcia, 2008) in which mined knowledge is integrated into the loop of the system not only to facilitate and enhance learning as a whole, but also to filter mined knowledge for decision making (Romero et al., 2008) or even to create intelligence upon which students, instructors, or administrators can build, can notably change academic behaviour (Baepler & Murdoch, 2010).

From an administrative perspective, Chang (2006) argues that the predictive capacity of data mining can further enhance enrolment management strategies by increasing the HEIs' understanding about their admitted applicants. Similarly, Delavaria, Phon-Amnuaisuka, and Reza Beikzadehb (2008) contend that data mining knowledge techniques are capable of enabling higher learning institutions to make better decisions, to put more advanced planning into place to direct students, and to predict individual behaviours with higher accuracy, and, in so doing, to enable the institutions to allocate resources and staff more effectively. Without inflating the merits of data mining in rethinking administrative and academic processes, it is clear that data-mining is gaining ground and is providing powerful analytical tools capable of converting untapped LMS and EPR data into critical decision-making tools with the potential of enhancing students' learning experiences (Garcia et al., 2011).

From a learning perspective, according to Castro, Vellido, Nebot, and Mugica (2007), data mining is being used in higher education

- to assess students' learning performance
- to provide feedback and adapt learning recommendations based on students' learning behaviours
- to evaluate learning materials and web-based courses, and
- to detect atypical students' learning behaviours.

Following this line of thinking, Perera, Kay, Koprinska, Yacef, and Zaiane (2009) used clustered data mining techniques to support the learning of group skills by building automated mirroring tools capable of facilitating group work. In a similar study, Sun, Cheng, Lin, and Wang (2008) used rules based on data mining results to form high interaction-learning groups.

For their part, Hung and Zhang (2008) applied data mining techniques to server logs, both to reveal online learning behaviour patterns and to support online learning management, facilitation, and design. Their study's results revealed students' behavioural patterns and preferences, which helped them to identify active and passive learners and which extracted important parameters for the prediction of the students' performance (Hung & Zhang). Using a similar approach, Ba-Omar, Petrounias, and Anwar (2007) analysed web access logs to identify learning patterns and offline learning styles. In a recent study, Abdous and He (2011) used text mining as a detection tool for the common technical problems faced by students taking video streaming courses.

Elsewhere, Zaiane and Luo (2001) analysed server logs to understand online learners' behaviours in an effort to improve their web-based learning environments. Later, Zaiane (2002) used association rule mining to construct a recommender-system based on data from online learners' profiles, access histories, and collective navigation patterns. This system can "intelligently" recommend learning activities or shortcuts to learners, based on the actions of previous learners. Similarly, Burr and Spennemann (2004) have pointed out that analysis of the patterns of user behaviour is important from both the technical and the pedagogical perspectives in order to predict network and traffic load, to align pedagogy with users' behaviours, and to plan and deliver services in a timely manner.

For their part, Dringus and Ellis (2005) proposed a data mining approach for "discovering and building alternative representations for the data underlying asynchronous discussion forums." This approach is intended to improve the instructor's ability to evaluate the progress of a threaded discussion. More recently, Lin, Hsieh, and Chuang (2009) conducted a study to investigate the potential of an automatic genre classification system (GCS) that can be used to facilitate the coding process of the content analysis of a threaded discussion forum,

Of particular relevance to our study, we discovered several studies which have used various EDM techniques to predict students' performance as measured by final grades. Minaei-Bidgoli and Punch (2003) used web-use features such correct answers, number of attempts for doing homework, total time spent on problems, participation in communication, and reading of material as predictors of students' final grades. Their prediction accuracy varied between 51% and 86.8%, depending on the type of classifier used. Similarly, Falakmasir & Jafar (2010) used data mining to rank students' activities which affected their performance, as measured by their final grade. Their findings suggest that students' participation in virtual classrooms had the greatest impact on their final grades.

For their part, Zafra and Ventura (2009) used a grammar-guided genetic programming algorithm to predict students' success or failure. These predictions were used to provide alternative learning activities that would enhance the students' chances of success.

Using Learning Management Systems-generated student tracking data (Macfadyen & Dawson, 2010), we propose the development of a customizable dashboard-like reporting tool. This tool is intended to provide instructors with real-time data on both students' engagement and the likelihood of their success. Unsurprisingly, their findings confirm that students' contribution to the course discussion board is the strongest predictor of their success.

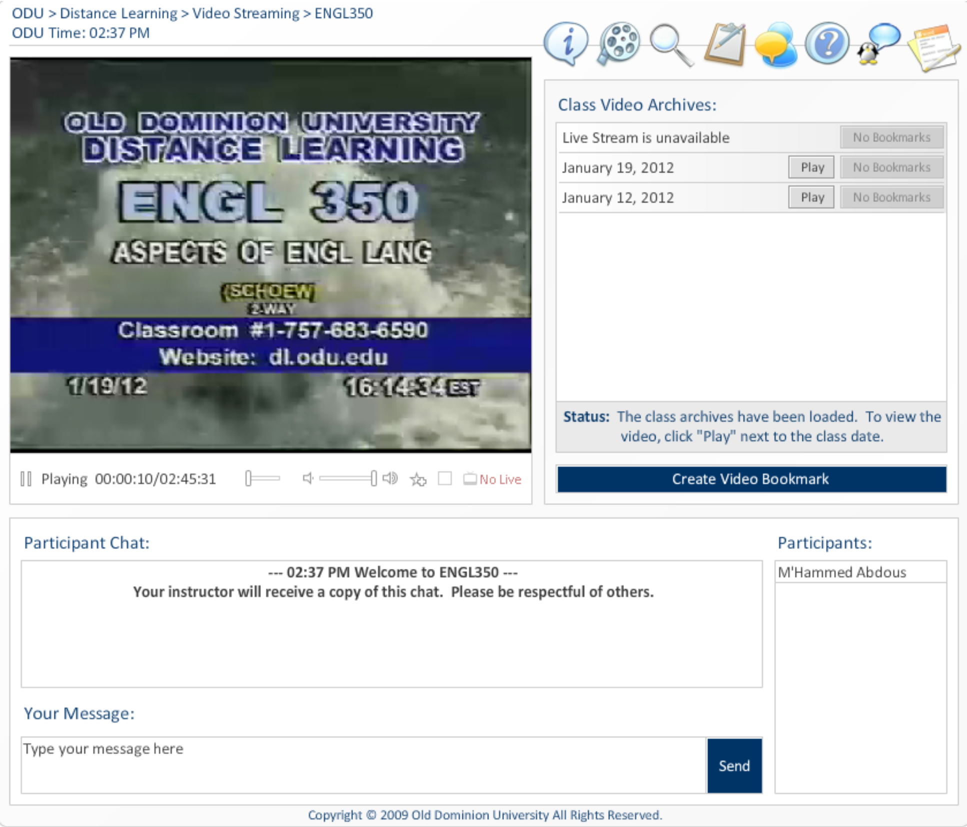
In reviewing the literature, Romero, Espejo, Zafra, Romero, and Ventura (2010) identified several avenues for using classification in educational settings: discovering student groups with similar characteristics, identifying learners with low motivations, proposing remedial actions, and predicting and classifying students using intelligent tutoring systems.

For their parts, Anand Kumar & Uma (2009) used the classification process to examine various attributes affecting student performance. Castellano and Martínez (2008) used collaborative filtering techniques to exploit students' grades in order to generate group profiles which could facilitate academic orientation. Along the same lines, Vialardi et al. (2011) used data mining techniques which employed the students' academic performance records to design a recommender system in support of the enrolment process.

In sum, this quick overview of the literature suggests that using various data mining techniques to predict students' performance as measured by final grades has been examined by several different studies of traditional learning management systems. However, none of the studies has explored the dynamics of online interaction in a live video streaming environment.

With these considerations in mind, we aim to apply both regression analysis and clustering analysis in order to explore students' learning behaviours (students' participation, login frequency, number of chat messages, and the type of questions submitted to instructor) along with their final grades. More specifically, we attempt to answer the following two questions:

- What are the major themes emerging from LVS students' online questions?
- How do these emerging themes predict the students' course grades?



ODU > Distance Learning > Video Streaming > ENGL350
ODU Time: 02:37 PM

Class Video Archives:

Live Stream is unavailable		No Bookmarks
January 19, 2012	Play	No Bookmarks
January 12, 2012	Play	No Bookmarks

Status: The class archives have been loaded. To view the video, click "Play" next to the class date.

--- 02:37 PM Welcome to ENGL350 ---
Your instructor will receive a copy of this chat. Please be respectful of others.

Copyright © 2009 Old Dominion University All Rights Reserved.

Figure 1. LVS interface

Methodology

Context of the study

This study was conducted in a public research university in the mid-Atlantic region which serves 17,000 undergraduate and 6,000 graduate students and offers more than 70 bachelor's degree programs, 60 master's degree programs, and 35 doctoral degree programs in a variety of fields. Located in a major maritime, military, and commerce hub, this institution offers strong emphases in science, engineering, and technology, especially in the maritime and aerospace sciences. 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 twenty-five years. This extended distance learning capability provides the university with a variety of delivery mode options (i.e., ways in which a course can be delivered). Courses can be offered simultaneously via three different delivery formats: face-to-face, via satellite broadcasting, and via live video-streaming. 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 during which they can interact with their instructor by sending text messages through the LVS course interface. 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 and without overburdening instructors during their class time (Figure 1).

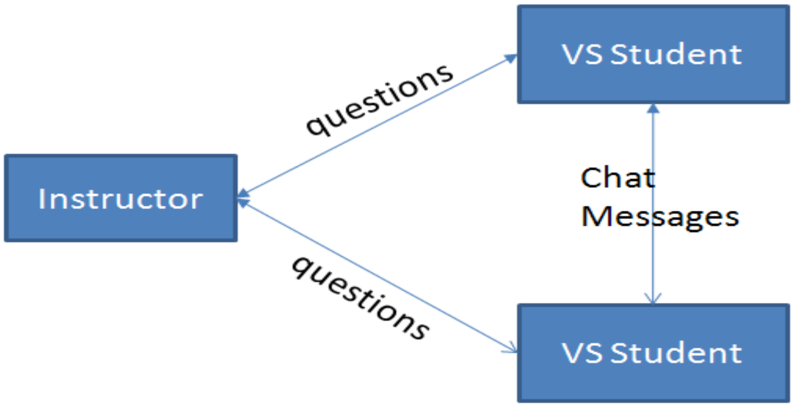


Figure 2. Interaction in VS courses

Participants

In total, 1,144 students completed 138 courses in a variety of subjects (e.g., accounting, computer engineering, information technology, human services, etc.) via the video streaming (VS) delivery mode during the Fall semester of 2009. All of the student-to-instructor questions, the two-way student-to-student chat messages, and the total login times were collected. Those VS students who never asked questions or chatted with their peers online were excluded from the actual analysis. The reasons why those students failed to get involved in the VS course discussion are suggested to be included in future investigation. One possible explanation is that some instructors never took the effort to invite their VS students to ask questions or to engage in online discussion. As a result of the pre-processing of the data, 298 students (those with complete information about their number of questions, number of chat messages, total login times, and final grade) were included in the data analysis. Due to factors such as privacy and university policy, the university's registrar's office could not provide us with the age or gender of these students, nor could we obtain the grading scales of each course. (The grading scale for each course at our university is determined by that course's instructor.)

Table 1. Distribution of students by college

Colleges	Percentages
Art and Letters	114 students (38.5%)
Education	79 students (26.5%)
Engineering	75 students (25%)
Science	23 students (7.7%)
Undecided	7 students (2.3%)

Table 2. Distribution of students by academic level

Student academic level	Percentage
Undergraduate Students	138 students (46%)
Graduate Students	160 students (54%)

The questions and chat messages posted by those 298 students, along with their course ID, Student ID, the date, and a time stamp were saved in the database.

Clustering analysis: Identification of emerging themes

Our analytical approach included three phases: pre-processing, in which raw data is transformed into a usable format, mainly by cleaning, assigning attributes, and integrating data; mining the data by applying various mining strategies and tools such as classification, clustering, and visualization; and post-processing, which allows for interpretation and use of the gained knowledge in rethinking processes or in making decisions (Garcia et al., 2011).

All of the questions posted by the students were recorded in the Microsoft SQL server Database. To prepare the data processing for clustering analysis, we wrote a program using the PHP programming language to aggregate questions from the same students within the same course in order to form a case which included the sequence of questions posted by the students.

Subsequently, we used NVivo 9 software to apply an automatic coding technique to each of the student question cases. Nvivo is a leading qualitative analysis tool on the market and has been used and tested by many researchers for content analysis (Zha, Kelly, Park & Fitzgerald, 2006). Automated coding one of NVivo 9's features; it allows for automatic coding of a text document by text strings. After nodes were generated from each student question case, a clustering analysis was conducted in order to classify these nodes into different clusters with NVivo 9. According to Nvivo, nodes are containers for specific themes, people, places, organizations, or other areas of interest.

Researchers can organize nodes into hierarchies – moving from more general topics (the parent node) to more specific topics (child nodes) – in order to support their particular research needs. Clustering analysis is a well-studied technique in data mining (Lin, et al., 2009) that uses an exploratory technique to visualize patterns by grouping sources which share similar words or attribute values, or which are coded similarly. From a data mining perspective, clustering is the unsupervised discovery of a hidden data concept. This approach is used in those situations in which a training set of pre-classified records is unavailable. In other words, this technique has the advantage of uncovering unanticipated trends, correlations, or patterns; no assumptions are made about the structure of the data (Chen & Liu, 2004)

The purpose of clustering analysis in this study is to classify students based on the student-shared characteristics in their questions. The cluster analysis tool in the NVivo 9 software confers upon researchers a different perspective on the unstructured textual data. Using the calculated similarity in each word that appears in the text of the nodes, NVivo 9 groups the nodes into a number of clusters. In our study, a statistical method named the Pearson correlation coefficient (-1 = least similar, 1 = most similar) was used as the similarity metric for the clustering analysis. The Pearson correlation coefficient is the preferred similarity metric used with Nvivo. More information about the clustering analysis of Nvivo can be found in Nvivo's online documentation website, http://www.qsrinternational.com/support.aspx.

To gain further insight from the textual questions or chat messages, we also applied the SPSS Clementine tool, which allowed us to analyse the unstructured textual data. The SPSS Clementine tool provides linguistic methods (extracting, grouping, indexing, etc.) for researchers to use in order to explore and extract key concepts from the text. As the result of the text mining, key concepts in our study were extracted and identified for analysis.

Measurement of final grade

The students' final grades, submitted to the University Registrar by each course instructor, were supplied to us by the University Registrar. In the actual data analysis, the final grades were categorized into three groups: A- to A, B- to B+, and Others.

Quantitative data analysis: Predictive relationship between online question theme and final grade

In the current study, all of the quantitative data analysis was implemented using SPSS 17.0. Furthermore, the alpha levels were set at the .05 level for all significance tests.

Due to the ordinal nature of the final grade, ordinal logistic regression analysis (Norusis, 2008; O'Connell, 2006) was implemented in order to examine the predictive relationship between the online question theme as the predictor and the final grade as the criterion variable. Specifically, a cumulative odds model was fitted to the data. The use of ordinal logistic regression, which was closely related to logistic regression, helped to avoid the statistical consequences that could occur from the violation of assumptions in linear regression, such as normality of errors and linearity in the parameters (King, 2008). The log transformation in logistic regression also ensured that the predicted probabilities for the event of interest would range from 0 to 1 without imposing the numerical constraint on the predicted log odds from the logistic equation (Cohen, Cohen, West, & Aiken, 2003). Given the ordinal nature of the final grade, ordinal logistic regression was used to take into account the information regarding the rank ordering of the outcomes (Hosmer & Lemeshow, 2000).

The overall predictive utility of the ordinal logistic model with the online question theme as the predictor was assessed by testing the improvement of the model fit relative to the null model with no predictor, with the χ^2 likelihood ratio test of the differences in deviances (O'Connell, 2006). The individual parameter estimate (i.e., the location coefficient) for the predictor variable was tested with the Wald test (Norusis, 2008). In ordinal logistic regression, two cutoffs (A- and B-) were used sequentially to form the cumulative odds equal to or higher than those two cutoffs, respectively. As a result, the probabilities of falling into three possible categories of final grade (A- to A; B- to B+; and Others) could be derived. Two different pseudo R^2 (Cox and Snell R^2 and Nagelkerke $R^{2)}$ were also computed in order to quantify the overall model fit (O'Connell). The larger the pseudo R^2, the better the model fit.

The parallel lines assumption in ordinal logistic regression was checked with the χ^2 likelihood ratio test (Norusis, 2008) to see if the relationship between those two research variables remained the same across two cutoffs (A- and B-).

Results

Identification of online question themes

In the current study, questions from each student during a semester were combined into one student entry so that students could be classified into different clusters based on the characteristics of their questions. The cluster analysis tool calculated each different word that appeared in the text of the entries by using the similarity metric. Then the entries were grouped into a number of clusters by NVivo 9, based on the calculated similarity index between each pair of entries. As a result, four major clusters of students were formed, based on the similarity of their questions. A multi-level, multiple cluster hierarchical structure was generated by clustering analysis (see Figure 3). These clusters were reviewed and interpreted collectively by two researchers and a graduate assistant. The two researchers had recently received specialized training about Nvivo 9 from the software producer. Differences in the review were compared, discussed, and resolved to reach an agreement. The coding results were further reviewed and discussed with an educational researcher to validate their accuracy.

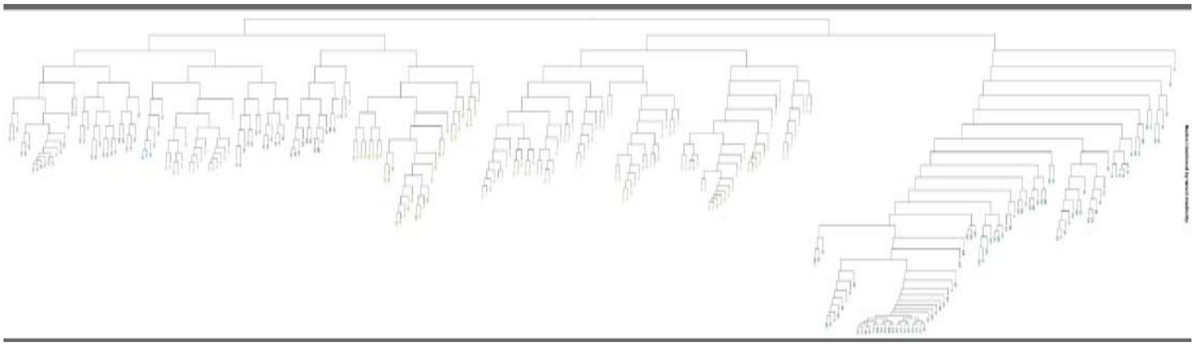


Figure 3. A cluster hierarchical structure

After a close review of the student questions in each cluster, four major themes were found (see Table 3).

Table 3. Four major themes in students' online questions

Cluster	***N***	***Theme***	***Text Content***
1	90	Check-in	Class check-in
2	87	Deadline/Schedule	Submission deadline, exam schedule, lab schedule
3	70	Evaluation/Technical	Exam format, grading, office hours, and technical problems
4	51	Learning/Comprehension	Questions regarding course materials and assignments

Descriptive statistics of final grade

The descriptive statistics of students' final grades by their online questions are listed in Table 4. Overall, about half (144, 48.32%) of the participants obtained a grade of A- or higher. Among the 298 participants, 90 posted mainly check-in questions and 87 posted questions related to deadline and schedule. The number of participants who posted questions mostly related to learning and comprehension was the lowest, relative to the number of their counterparts posting questions on other themes.

Table 4. Descriptive statistics of final grade by online question theme (N = 298)

	Online question theme									
	1		2		3		4		Total	
Final Grade	n	%	n	%	n	%	n	%	n	%
A- to A	43	47.78	36	41.38	32	45.71	33	64.71	144	48.32
B- to B+	28	31.11	29	33.33	23	32.86	15	29.41	95	31.88
Others	19	21.11	22	25.29	15	21.43	3	5.88	59	19.80
Total	90	100.00	87	100.00	70	100.00	51	100.00	298	100.00

Note. Question theme 1: Check-in; Question theme 2: Deadline/Schedule; Question theme 3: Evaluation/Technical; Question theme 4: Learning/Comprehension.

Predictive relationship between online question theme and final grade

In the ordinal logistic regression model (see Table 5), the results of the chi-square likelihood ratio test supported a nonzero predictive relationship between the online question theme and the final grade, χ^2 (3, N = 298) = 10.017, p < .05. Furthermore, the results did not indicate the violation of the parallel lines assumption, χ^2 (3, N = 298) = 2.051, p > .05. Therefore, the predictive relationship between the online question theme and the final grade remained constant across two cutoffs of final grade (Norusis, 2008). The Cox and Snell R^2 and the Nagelkerke R^2 were .033 and .038 respectively, and indicated a modest predictive relationship. Overall, the online question theme would prove to be a useful predictor for the final grade.

The logistic regression coefficients (i.e., the location coefficients) for question themes 1, 2, and 3 were all positive and were statistically significant at the .05 level. Due to the way in which the ordinal logistic regression model was set up in SPSS (Norusis, 2008), the above statistically nonzero, positive regression coefficients suggested that the odds of getting a higher final grade, relative to all lower final grades at various cutoff values, were higher for the participants whose questions concerned learning/comprehension (Theme 4) in comparison with participants with the other three question themes (i.e., 1: Check-in; 2: Deadline/Schedule; 3: Evaluation/Technical). Specifically, for participants with the question theme of Learning/Comprehension, the odds of obtaining a grade equal to or higher than those two cutoffs (A- and B-), relative to all other lower grades, were 2.214 times higher than for the students whose questions had the theme of check-in, 3.020 times higher than those whose questions had the theme of deadline/schedule, and 2.361 times higher than the students whose questions had the theme of evaluation/technical. While using Question Theme 1, or Question Theme 2, or Question Theme 3 as the reference category respectively, no differences in the odds of obtaining better grades were found among the three theme groups.

The computed predicted probabilities of obtaining a final grade of A- to A+, B- to B+, and Others, respectively, for participants in those four question theme groups (1: Check-in; 2: Deadline/Schedule; 3: Evaluation/Technical; 4: Learning/Comprehension), were 47.07%, 32.89%, 20.04% in the Theme 1 group, 40.75%, 34.77%, 24.48% in the Theme 2 group, 45.48%, 33.43%, 21.09% in the Theme 3 group, and 66.31%, 23.51%, 10.17% in the Theme 4 group.

Table 5. Ordinal logistic model with online question theme as the predictor for final grade (N =298)

Parameter	Estimate	*Wald*
Location		
Question Theme 1	.795*	5.078
Question Theme 2	1.052*	8.834
Question Theme 3	.859*	5.452
Threshold		
Grade = A- to A	.677*	5.394
Grade = B- to B+	2.178*	47.867

Overall model evaluation	χ^2	*df*	*Cox and Snell* R^2	*Nagelkerke* R^2
Likelihood ratio test	*10.017**	*3*		
Goodness-of-fit index			*0.33*	*.38*

Note. Question Theme 1: Check-in; Question Theme 2: Deadline/Schedule; Question Theme 3: Evaluation/Technical; Question Theme 4: Learning/Comprehension as the reference category.
$*p < .05$.

Two cutoffs were set for the ordinal criterion variable, final grade, to examine how the increase in the faculty engagement score was related to the change in the odds, and in turn, to the probability of obtaining a higher final grade (O'Connell, 2006). The odds of obtaining a higher final grade at two cutoffs were the ratios of the probabilities of: A to all lower grades, and A through B- to all lower grades. The faculty engagement scores as the sample mean (i.e., 20.697 in raw score) and the one standard deviation (i.e., 5.560 in raw score) above the sample mean were examined to demonstrate the way in which the probability of obtaining a higher course final grade changed with the increase in faculty engagement (Norusis, 2008). Given an increase of one standard deviation in the faculty engagement score from the sample mean (i.e., from 20.697 to 26.257), the predicted probability of obtaining a final grade of A increased from 46.71% to 59.78% at the first cut-off. At the second cut-off, the predicted probability of obtaining a final grade of B- or higher increased from 78.79% to 86.30%.

Moreover, with the faculty engagement score as the sample mean (i.e., 20.697 in raw score), the predicted probabilities of obtaining one of those three categories of course final grade (A, A- to B-, or Other) were 46.71%, 32.08%, and 21.21% respectively. While the raw faculty engagement score increased by one standard deviation to 26.257, the predicted probabilities of obtaining one of those final grades became 59.78%, 26.52%, and 13.70% respectively. Therefore, the increase in the faculty engagement score was accompanied by the increased probability of obtaining a better course final grade.

Discussion

A student's final grade depends on many factors, including the student's motivation, learning style, and previous background, the instructor's teaching and grading scales, the exam's and assignment's difficulty levels, etc. A holistic view of student demographic and institutional variables, as opposed to the single variable, must be examined in determining the overall online learning experience (Herbert, 2008).

In this study, our data shows that online VS student participation cannot be safely used to predict final grades. Perhaps the uniqueness of our VS interface (text-based chat in a live-video-streaming environment) explains our findings. Otherwise, previous studies including Macfadyen and Dawson's study (2010) found that students' participation and contribution to discussion boards in traditional learning management systems remain some of the strongest predictors of students' success.

However, our analysis found that there is a correlation between questions posed to instructors and chat messages posted among students. Those who chat often also interact more often with their instructor.

We also analysed the chat messages (student-to-student communications) using the SPSS Clementine text mining tool. We noticed two outstanding concepts in the students' chat messages (among themselves) and their frequency: they discussed technical problems (videos, sound, etc.) at 5% and test/exam issues at 2%. However, they addressed the same concepts in their messages to the instructor with this frequency: technical problems at 2% and test/exam issues at 2%. Thus, it seems that students are more likely to discuss technology problems with their peers and try to help each other than to discuss those issues with their instructors.

The messages also revealed interaction patterns including topics related to project and assignment collaboration, discussion of grades, socialization, and greetings. In addition, the data reveals that students with a higher number of logins asked more questions and exchanged more chat messages with their classmates. In contrast, students with fewer logins rarely participated in the class; in fact, some of them rarely even logged into the system.

Conclusions and future research

This study was conducted in order to exploit the untapped data generated by LVS students. Our results revealed several student learning behaviours, ranging from active participation and interaction with the instructor to a lack of participation or even of attendance. Overall, our findings corroborate those of a previous study (Abdous & He, 2011). In spite of the limitations related to self-selection bias and to the use of final grades as a measurement of student learning outcomes (Abdous & Yen, 2010), we believe that we can provide some ways in which the learning experiences of LVS students can be improved and made more successful, based on our years of experience of working with faculty who teach VS courses. To this end, the following recommendations are made:

- Ensure faculty readiness and training prior to teaching LVS courses.
- Develop facilitation techniques to assist faculty in integrating LVS students into the dynamics of the classroom.
- Implement a tracking system for LVS students' attendance.
- Encourage active participation and interaction during LVS sessions.
- Provide students with tips on effective participation and interaction during LVS sessions (writing messages, timing of questions, etc.)

As we make these recommendations, we reiterate that educational data-mining is clearly providing powerful analytical tools capable of converting untapped LMS and EPR data into critical decision-making information which has the capability of enhancing students' learning experiences (Garcia et al., 2011). While adding to the body of literature, our hybrid approach provides a solid framework that can be used to exploit educational data to rethink and improve the learning experiences of students using some of the various new delivery modes that are currently reshaping higher education. Further understanding of students' engagement and the dynamics of their interaction in and with these new delivery modes will contribute to the promulgation of an effective and engaging learning experience for all.

References

Abdous, M., & He, W. (2011). Using text mining to uncover students' technology-related problems in live video streaming. *British Journal of Educational Technology*, *42*(1), 40-49.

Abdous, M., & Yen, C.-J. (2010). A predictive study of learner satisfaction and outcomes in face-to-face, satellite broadcast, and live video-streaming learning environments. *The Internet and Higher Education*, *13*, 248-257.

Anand Kumar N, & Uma, G. (2009). Improving academic performance of students by applying data mining technique. *EuroJournals, 34*, pp. 526-534.

Baepler, P., & Murdoch, C. J. (2010). Academic analytics and data mining in higher education. *International Journal for the Scholarship of Teaching & Learning, 4*(2), 1-9.

Baker, R., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions 2009. *Journal of Education Data Mining, 1*(1), 3-17. Retrieved from http://www.educationaldatamining.org/JEDM/images/articles/vol1/issue1/JEDMVol1Issue1_BakerYacef.pdf

Ba-Omar, H., Petrounias, I., & Anwar, F. (2007). A framework for using web usage mining to personalise e-learning. In M. Spector et al. (Eds.), *Proceedings of the Seventh IEEE International Conference on Advanced Learning Technologies. ICALT 2007* (pp.937-938). Los Alamitos, California: IEEE Computer Society Press

Black, E., Dawson, K., & Priem, J. (2008). Data for free: Using LMS activity logs to measure community in an online course. *Internet and Higher Education, 11*(2), 65-70.

Burr, L., & Spennemann, D. H. R. (2004). Patterns of user behaviour in university online forums. *International Journal of Instructional Technology and Distance Learning, 1*(10), 11-28.

Castellano, E., & Martínez, L. (2008, July). ***ORIEB, A CRS for academic orientation using qualitative assessments.*** Paper presented at the IADIS International Conference E-Learning, Amsterdam, The Netherlands.

Castro, F., Vellido, A., Nebot, A., & Mugica, F. (2007). Applying data mining techniques to e-learning problems. In L. C. Jain, R. Tedman, & D. Tedman (Eds.), *Evolution of Teaching and Learning Paradigms in Intelligent Environment* (pp. 183-221). New York: Springer-Verlag.

Chang, L. (2006). Applying data mining to predict college admissions yield: A case study. *New Directions for Institutional Research, 2006*(131), 53-68.

Chen, S. Y., & Liu, X. (2004). The contribution of data mining to information science. *Journal of Information Science, 30*(6), 550-558.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Delavaria, N., Phon-Amnuaisuka, S., & Reza Beikzadehb, M. (2008). Data mining application in higher learning institutions. *Informatics in Education, 7*(1), 31.

Dringus, L. P., & Ellis, T. (2005). Using data mining as a strategy for assessing asynchronous discussion forums. *Computers and Education, 45*(1), 141-160.

Falakmasir, M., & Jafar, H. (2010, June). *Using educational data mining methods to study the impact of virtual classroom in e-learning.* Paper presented at the Proceedings of the 3rd International Conference on Educational Data Mining, Pittsburgh, PA, USA.

Faulkner, R., Davidson, J. W., & McPherson, G. E. (2010). The value of data mining in music education research and some findings from its application to a study of instrumental learning during childhood. *International Journal of Music Education, 28*(3), 212-30.

Garcia, E., Romero, C., Ventura, S., & de Castro, C. (2011). A collaborative educational association rule mining tool. *Internet and Higher Education, 14*(2), 77-88.

Harding, J. A., Shahbaz, M., Srinivas, M., and Kusiak, A. (2006). Data mining in manufacturing: A review. *Journal of Manufacturing Science and Engineering, 128*(4), 969–976.

Hen, L. E., & Lee, S. P. (2008). Performance analysis of data mining tools cumulating with a proposed data mining middleware. *Journal of Computer Science, 4*(10), 826-833.

Herbert, M. (2008). Staying the course: A study in online student satisfaction and retention. *Online Journal of Distance Learning Administration,* 9(4). Retrieved from http://www.westga.edu/~distance/ojdla/winter94/herbert94.htm

Hosmer, D.W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.

Hung, J., & Zhang, K. (2008). Revealing online learning behaviors and activity patterns and making predictions with data mining techniques in online teaching. *MERLOT Journal of Online Learning and Teaching, 4*(4). Retrieved from http://jolt.merlot.org/vol4no4/hung_1208.htm

Hung, J.-L., & Crooks, S. M. (2009). Examining online learning patterns with data mining techniques in peer-moderated and teacher-moderated courses. *Journal of Educational Computing Research, 40*(2), 183-210.

King, J. E. (2008). Binary logistic regression. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 358-384). Thousand Oaks, CA: Sage Publications.

Lin, F.-R., Hsieh, L.-S., & Chuang, F.-T. (2009). Discovering genres of online discussion threads via text mining. *Computers & Education, 52*(2), 481-495.

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education*, *54*(2), 588-599.

Minaei-Bidgoli, B., Kashy, D. A., Kortemeyer, G., & Punch, W. (2003, November). Predicting student performance: An application of data mining methods with an educational web-based system (LON-CAPA). Paper presented at the 33rd ASEE/IEEE Frontiers in Education Conference, Boulder, CO, USA. Retrieved from http://lon-capa.org/papers/v5-FIE-paper.pdf

Ngai, E. W. T., Xiu, L., and Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature relationship and classification. *Expert Systems with Applications, 36*, 2529-2602.

Norusis, M. (2008). ***SPSS statistics 17.0 advanced statistical procedures companion.*** Upper Saddle River, NJ: Prentice Hall.

O’Connell, A. (2006). *Logistic regression models for ordinal response variables*. Thousand Oaks, CA: Sage Publications.

Perera, D., Kay, J., Koprinska, I., Yacef, K., & Zaiane, O. (2009). Clustering and sequential data mining of online collaborative learning data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(6), 759-772.

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transaction on Systems, Man, and Cybernetics, Part C: Applications and Reviews. 40*(6)*,* 601-618.

Romero, C., Espejo, P. G., Zafra, A., Romero, J. R., & Ventura, S. (2010). Web usage mining for predicting final marks of students that use Moodle courses. *Computer Applications in Engineering Education.* doi: 10.1002/cae.20456

Romero, C., Ventura, S., & Garcia, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education, 51*(1), 368-384.

Sun, P. C., Cheng, H. K., Lin, T. C., & Wang, F. S. (2008). A design to promote group learning in e-learning: experiences from the field. *Computers & Education*, *50*(3), 661–677.

Vialardi, C., Chue, J., Peche, J., Alvarado, G., Vinatea, B., Estrella, J., & Ortigosa, L. (2011). A data mining approach to guide students through the enrollment process based on academic performance. *User Modeling and User - Adapted Interaction, 21*(1-2), 217.

Zafra, A., & Ventura, S. (2009, June). *Predicting student grades in learning management systems with multiple instance programming.* Paper presented at the Proceedings of the 2nd International Conference on Educational Data Mining, Cordoba, Spain.

Zaïane, O. R. (2002). Building a recommender agent for e-learning systems. In L. Aroyo et al. (Eds.), *Proceedings of the International Conference on Computers in Education ICCE '02* (55-59). Washington, DC: IEEE Computer Society Press

Zaiane, O. R., & Luo, J. (2001). Towards evaluating learners' behavior in a web-based distance learning environment. Retrieved from http://webdocs.cs.ualberta.ca/~zaiane/postscript/icalt.pdf

Zha, S., Kelly, P., Park, M. K., & Fitzgerald, G. (2006). An investigation of ESL students using electronic discussion boards. *Journal of Research on Technology in Education, 38*(3), 349-367.