Identifying Latent Study Habits by Mining Learner Behavior Patterns in Massive Open Online Courses

Miaomiao Wen
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
mwen@cs.cmu.edu

Carolyn Penstein Rosé
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
cprose@cs.cmu.edu

ABSTRACT

MOOCs attract diverse users with varying habits. Identifying those patterns through clickstream analysis could enable more effective personalized support for student information seeking and learning in that online context. We propose a novel method to characterize types of sessions in MOOCs by mining the habitual behaviors of students within individual sessions. We model learning sessions as a distribution of activities and activity sequences with a topical N-gram model. The representation offers insights into what groupings of habitual student behaviors are associated with higher or lower success in the course. We also investigate how context information, such as time of day or a user's demographic information, is associated with the types of learning sessions.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications

Keywords

Massive Open Online Course; MOOCs; learning behavior patterns; sequence mining

1. INTRODUCTION

Massive Open Online Courses (MOOCs) have recently garnered widespread public attention for their potential to make high quality education accessible to everyone. Researchers, educators, and the general public have become interested in understanding learning experiences in MOOCs. A major component of the learning experience is navigation through course content. The MOOC literature so far has focused on a summative view of user participation over extended periods of time, such as over the whole course or over a whole week - trying to identify different groups of users based on aggregated statistics such as total number of lectures and assignments completed by the student [1]. We develop a more fine grained representation of MOOC clickstream data using topical N-gram models with single sessions as the unit of analysis, with the goal of identifying habbits and strategies.

CIKM'14, November 3–7, 2014, Shanghai, China.

ACM 978-1-4503-2598-1/14/11 ...$15.00.
http://dx.doi.org/10.1145/2661829.2662033.

MOOC	#Users	#Sessions	#Clicks
Virtual Instruction	20,372	76,770	2,112,857
Financial Planning	112,318	469,238	10,073,997

Table 1: Statistics of the two Coursera MOOCs.

In a MOOC, each student's complete interaction with the course materials is recorded as a clickstream, similar to other online environments. However, the different nature of the interaction provides new challenges for the community of researchers developing techniques for clickstream analysis. Understanding how students interact with MOOCs is a crucial issue because it affects how MOOC engineers and instructors design future online courses. Effective analysis of clickstream data could enable dynamic interventions to improve success of students to achieve their information access goals. What makes the mining of learning behavior patterns hard is that learning sessions are typically composed of a range of loosely coordinated activities. The composition of a session is highly variable depending on factors such as learner engagement and time available. Previous behavior mining methods, such as those designed for search engine clicklog analysis, are not quite rich enough to capture these distinctions. We propose to leverage topical N-gram models to (1) capture the typical combination of learning activities and (2) identify frequent learning activity sequence which can suggest learning strategies. Thus we can characterize each session of MOOC interaction as a composition of such activities or activity sequence patterns. We find that there are typically a small number of latent session types across MOOCs we have studied. In this paper we present results that demonstrate that we can observe meaningful differences in student orientations towards course content. To understand how the learning sessions can be affected by context or individual user's engagement level, we mine the associations between user session type and context information captured by system trace logs.

2. DATASET AND PRE-PROCESSING

Our dataset consists of two Coursera[1] MOOCs, one is about virtual instruction(VI) and the other one is about personal financial planning(FP). They were both offered in 2013. Table 1 shows an overview of the data we extracted from the system trace logs.

[1]https://www.coursera.org/

MOOC	None ≤ 10%	Fail 10%-60%	Pass 60%-90%	Distinction ≥90%
VI	17,562	1,418	507	881
FP	71,522	8,768	1,025	2,918

Table 2: Number of students within each final grade level.

Pre-processing is the first part of Web Usage Mining which includes the domain dependent tasks such as data cleaning and session identification. In our case, a new session is considered to begin when the time interval between two successive inter-transaction clicks adds up to 60 minutes. The visited or submitted contents during each session are considered to be part of that session.

We extract 18 types of learning activities from the trace data. In particular, we distinguish between submitted quizzes, assignments, peer assessments and those that are attempted but not submitted. We also differentiate passive activities such as browsing and active activities such as publishing a post in the forums. Only the underlined activities are graded.

- Video lecture
  (1)Watch a video lecture (Lecture).
- Assignment
  (2)Browse an assignment (BrowseAssignment); (3)Submit an assignment (SubmitAssignment); (4)Start a peer-asse ssment[2] (BrowsePeerAssessment); (5)Submit a peer-assessment (SubmitPeerAssessment).
- Quiz
  (6)Submit an in-video quiz[3] (SubmitVideoQuiz); (7) Browse a weekly quiz (BrowseQuiz); (8)Submit a weekly quiz (SubmitQuiz). (9)Browse the final quiz (BrowseFinalQuiz); (10)Submit a final quiz (SubmitFinalQuiz).
- Survey
  (11)Browse a pre-course or post-course survey (BrowseSurvey); (12)Submit a survey (SubmitSurvey).
- Forum participation
  (13)Browse a thread in the course forums (BrowseForum); (14)Post a new post in the course forums (Post); (15)Comment on the other posts in the course forums (Comment); (16)Upvote a post(Upvote); (17)Downvote a post (Downvote).
- Browse the course material
  (18)Browse the course material without clicking on videos, quizzes, surveys, forums or assignments (BrowseCourse).

We extract information for each session and each user. The contextual information associated with each session includes: (1) Time information, including *Hour of the day*, *Day Period*, *Day of the week* and *Course week*. (2) *Device* used during the current session. The device can be Desktop, Tablet or Mobile. (3) *Length* of the session. A Short session is shorter than 5 minutes. A Long session is longer than 30 minutes. Otherwise the session length is Medium.

Information extracted for each user includes: *Gender*, *Age*, *Country of Origin* and *Final Grade*. *Gender* and *Age* are only known for the users who filled out the pre-course survey. *Country of Origin* is determined based on the IP address. As there are more than 100 different countries of origin, we group them into *US* and *NON-US*. As for achievement, we group the users into four groups according to their final grade. Students who successfully complete this class with a grade of around 60% - 70% will receive a Statement of Accomplishment. Students who complete this course with a grade of 90% or better will receive a Statement of Accomplishment with Distinction from Coursera. The statistics are shown in Table 2.

[2]Peer assessment is the practice of classmates evaluating each other's work.

[3]In-video quizzes are quizzes that pop up during the lectures.

3. METHODS

In this section, we describe how we model each learning session with a topical N-gram model and how we extract the typical context associated with different types of learning sessions.

3.1 Characterizing Learning Sessions with the Topical N-gram Model

A learning session can be characterized as a probabilistic combination of interaction and interaction sequence patterns. To motivate our work with intuitive examples, an intense learning session may be one that includes an assignment submission. In the same session, the user may watch video lectures to prepare for the assignment or go to the discussion forums to find related discussions. Just as word order and phrases are often critical to capturing the meaning of text in many text mining tasks, the order of activities within sessions is important for capturing a user's learning strategies. For example, a sequence, "FinalQuizBrowse_Lecture_FinalQuizSubmit", implies the user tries to refer to the lecture during the final quiz. Activity sequences as the whole may carry more information than an unordered collection of its individual components.

A topical N-gram model is a topic model that discovers topics as well as topical phrases [10]. The probabilistic model generates words, or actions in our case, in their order of appearance. This is accomplished by iterating over words, and for each word, first sampling a topic, then sampling its status as a unigram or bigram, and then sampling the word from the selected topic-specific unigram or bigram distribution. Successive bigrams can form longer phrases. Thus the model can distinguish that "white house" has a special meaning as a phrase in the 'politics' topic, but not in the 'real estate' topic. We apply topical N-gram models to our learning session modeling task where we treat each learning activity defined during pre-processing (Section 2) as a word and each session as a document. Then we can characterize learning sessions with topic proportions.

3.2 Mining Learner Behavior Patterns

Intuitively, some learner behaviors are context-sensitive, that is, the occurrences of these behaviors are influenced by contextual factors like time and course schedule. For example, some users prefer to have an intense learning session, which involves doing assignments, on Sundays but only browse the course forums on weeknights. The associations between user interaction records and the corresponding contexts, which can be referred to as behavior patterns, can be used to characterize user habits[2]. We are especially interested in how students' study habits influence their success. This is important because it may enable context sensitive support to be generated for students. Instead of traditional association rule mining, we use classification rule mining to

Session Topic	Top Activities and Activity Sequences
Lecture and Peer-Assessment	Lecture_Lecture, Lecture_Lecture_Quiz, Lecture, Quiz, SubmitPeerAssessment
Browse Course	BrowseCourse
Assignment and Forum	BrowseAssignment, BrowseForum, BrowseForum_BrowseForum, Lecture_Lecture, Upvote, SubmitQuiz, SubmitAssignment
Final Quiz and Survey	Quiz, BrowseForum_BrowseForum, BrowseSurvey, BrowseFinalQuiz, SubmitSurvey, BrowseLecture_BrowseLecture_BrowseFinalQuiz, SubmitFinalQuiz
Lecture and Quiz	Quiz, Lecture, Lecture_Lecture, Lecture_Lecture_SubmitVideoQuiz, SubmitVideoQuiz, Lecture_SubmitVideoQuiz, SubmitQuiz

Table 3: Topics generated by topical N-grams for the Virtual Instruction MOOC. Lecture_Lecture represents watching two different videos consecutively. BrowseForum_BrowseForum represents browsing two different forum threads consecutively

discover a small set of rules in the context logs that predict session types generated by topic modeling[7].

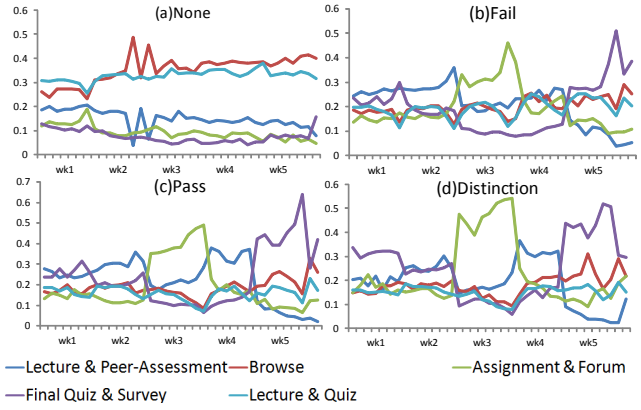

Figure 1: Daily average topic distributions for four final grade groups in the Virtual Instruction MOOC.

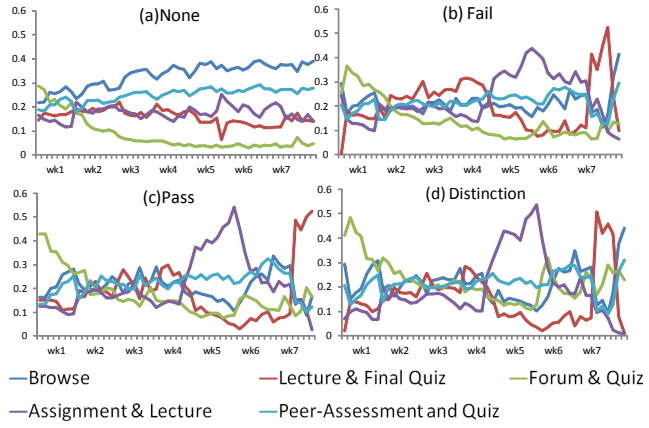

Figure 2: Daily average topic distributions for four final grade groups in the Financial Planning MOOC.

4. RESULTS

In this section, we first present the session topics generated from the topical N-gram model. The visualization of average topic distribution on each course day shows different engagement patterns of different final grade groups in our two MOOCs. Then we group each learning session based on its topic distributions. Interesting learning behavior patterns are mined from the context-rich log dataset.

4.1 Pattern Analysis

We show the top ranking activities and activity sequences for each topic for the VI MOOC in Table 3[4]. The number of topics is set to five. The topics generated by topical N-gram models capture both the typical combination of interactions and typical sequences of learning activities, such as "BrowseLecture_BrowseLecture_BrowseFinalQuiz", which implies the user watches more than one video lectures before taking the final quiz.

Based on the topic distribution, we assign a session type to each session with the largest topic proportion. We set the session type as the variable to predict and use classification rule mining to mine learner behavior patters. Our definitions of Confidence and Support are similar to those in [2]. On a course level, all the behavior patterns with Confidence larger than 0.25 are mined. Then we select at most the top 20 behavior patterns for each session type instead of using all the mined behavior patterns. We manually check the mined behavior patterns. Table 4 shows some patterns mined from the VI course. These behavior patterns reflect some interesting learning habits of the students in this course. For example, on course week 3, students are likely to engage in an *Assignment and Forum* session.

4.2 Validation

We now investigate how a student's final grade is related to her learning session distribution. Final grade can indicate student knowledge and also engagement. Here we think of the final grade as an independent variable; our goal is not to predict a student's grade from her activity but rather to gain insight into how high-grade and low-grade students distribute their activities differently along the course weeks. On each course day, we compute the average session topic proportions over all the learning sessions that happened on that day. In Figures 1 and 2 we show trends for our two MOOCs. We find that the distribution patterns are qualitatively similar. For example in both MOOCs, The None achievement users(Figure 1(a) and Figure 2(a)) are characterized with a flat distribution across the five learning session topics, which means they are insensitive to course schedules and deadlines. Since MOOCs have created space for these less performance-oriented types of learning, such as auditing or exploring a course, they have more purely *Browsing* sessions. The other three achievement groups reflect the course

[4]The activities for the FP MOOC are very similar but slightly different. To save space, we do not show them here.

Behavior Pattern	Support	Confidence
Context:{*Final Grade* = None, *Device* = Desktop, *Country* = NON-US, *Length* = Short} ⇒ *Session* = BrowseCourse	8,787	0.78
Context:{*Gender* = Male } ⇒ *Session* = BrowseCourse	3,888	0.49
Context:{ *Length* = Long } ⇒ *Session* = Lecture and Peer-Assessment	8,943	0.37
Context:{ *Course Week* = 3 } ⇒ *Session* = Assignment and Forum	4,048	0.34
Context:{ *Final Grade* = Distinction, *Device* = Desktop } ⇒ *Session* = Assignment and Forum	5,075	0.27
Context:{ *Final Grade* = Distinction, *Length* = Long } ⇒ *Session* = Final Quiz and Survey	5,521	0.26

Table 4: Top-ranking mined behavior patterns for the Virtual Instruction MOOC.

schedule in different levels. In the VI course (Figure 1), the assignment is released on Monday of course week 3, users have more *Assignment and Forum* sessions to check out or do the assignment. Towards the assignment deadline, which is the end of week 3, an even higher proportion of the sessions are *Assignment and Forum*. Similar trends (bumps in the curve) can also be observed for peer-assessment(released and due in course week 4) and final quiz(week 5). It is interesting to compare the trends between Fail and Pass(Figure 1(b) vs. Figure 1(c); Figure 2(b) vs. Figure 2(c)), they have similar trends except that Fail users mostly do not have the clear Peer-Assessment bump in week 4. This may largely be due to the fact that they did not finish their own assignments so they cannot do peer-assessment. If we compare Pass and Distinction users(Figure 1(c) vs. Figure 1(d); Figure 2(c) vs. Figure 2(d)), we can see that Pass users tend to "procrastinate" towards a deadline, as much more of their Assignment or Final Quiz related sessions are in the later part of the week (deadlines are on Sunday nights). We leave deeper analysis of these behavior patterns, such as procrastination, for future work.

5. RELATED WORK

Though there have been some quantitative, large-scale studies of student behavior in MOOCs to date, there is still much room for development of techniques that offer high resolution into student routines and habits at a fine grained level. Very few prior studies have utilized the full spectrum of rich information captured by activity trace data in an integrated way. Recently, Anderson et al.[1] have developed a taxonomy of individual learner behaviors related specifically to assignments, designed to examine the different behavior patterns aggregated across a student's entire experience in the course to distinguish high- and low-achieving students. Most commonly in prior studies, only one or two types or aspects of student interaction have been investigated at a time, for example, students navigating backwards [4], in-video dropouts[6], forum posting behaviors[11] and students' time on specific tasks [3].

Since late 1990, web usage mining has been widely studied. [9] surveyed the popular techniques in this field such as Association Rule Mining and Clustering. Sequential pattern discovery can characterize user episodes for the mining of traversal patterns on search engines, shopping sites, etc. A distinct property of the user interactions with MOOCs is that user navigation is less dependent on the linking structure in the website and more related to their course goals. Thus typical sequential pattern mining algorithms may be less suitable for our task. Inspired by [5], which uses topic models to discover patterns in a user's daily routine from sensor data, we adopt a form of topic model to extract leaning activity patterns.

6. CONCLUSION AND FUTURE WORK

In this paper, we propose a novel approach for characterizing learning behavior patterns. The experiments show that learning sessions can be modeled as combinations of several session topics, which provides insights into how high-grade and low-grade students distribute their activities differently along the course weeks. Based on context information associated with each learning session, we mine learning behavior patterns and then observe how these patterns play out over a course. In the future, we want to mine individual learning behavior patterns to discover latent learner types [8].

7. ACKNOWLEDGMENTS

This research was funded in part by NSF grants IIS-1320064 and OMA-0836012.

8. REFERENCES

[1] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec. Engaging with massive online courses. In *WWW'14*, pages 687–698, 2014.

[2] H. Cao, T. Bao, Q. Yang, E. Chen, and J. Tian. An effective approach for mining mobile user habits. In *CIKM'10*, 2010.

[3] J. Champaign, K. F. Colvin, A. Liu, C. Fredericks, D. Seaton, and D. E. Pritchard. Correlating skill and improvement in 2 moocs with a student's time on tasks. In *L@S'14*, pages 11–20, 2014.

[4] P. J. Guo and K. Reinecke. Demographic differences in how students navigate through moocs.

[5] T. Huynh, M. Fritz, and B. Schiele. Discovery of activity patterns using topic models. In *Ubicomp'08*, pages 10–19. ACM, 2008.

[6] J. Kim. Understanding in-video dropouts and interaction peaks inonline lecture videos. L@S '14, pages 31–40, 2014.

[7] B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In *KDD'98*, 1998.

[8] H. Ma, H. Cao, Q. Yang, E. Chen, and J. Tian. A habit mining approach for discovering similar mobile users. In *WWW'12*, pages 231–240. ACM, 2012.

[9] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. Web usage mining: Discovery and applications of usage patterns from web data. *ACM SIGKDD Explorations Newsletter*, 1(2):12–23, 2000.

[10] X. Wang, A. McCallum, and X. Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *ICDM 2007*, pages 697–702, 2007.

[11] M. Wen, D. Yang, and C. Rosé. Linguistic reflections of student engagement in massive open online courses. *In ICWSM*, 2014.