Text Analysis & Predictive Modeling for Smarter Online Education

Snigdha Chaturvedi
snigdhac@cs.umd.edu

Motivation

Developments in the past few years have fundamentally altered the traditional landscapes of pedagogy and teaching. Ubiquitous computing and high bandwidth internet connectivity in many parts of the world have reshaped the *modulus operandi* in distance education towards Massive Open Online Courses (MOOC). Courses offered by ventures such as Coursera and Udacity now impart inexpensive and high-quality education from field-experts to thousands of learners across geographic and cultural barriers. The social and economic implications of this archetype are momentous, leading the Time magazine to call free MOOCs the "Ivy League for the Masses"[1]. Even as the MOOC model presents exciting possibilities, it simultaneously presents a multitude of challenges that must first be negotiated to completely realize its potential.

Online platforms have been especially criticized on grounds of lacking personalization of the educational experience[2]. Identifying a student's level of interest and engagement could help modify teaching behaviour such as by altering teaching strategy, employing simpler examples in instruction, slowing or speeding up the pace of instruction, or switching to newer topics [11]. Due to the immense class sizes, MOOCS struggle with comprehending student needs at a personal level, changing course-structure according to student ability, and understanding class dynamics in terms of student interactions and communities.

Fortunately, from a computational perspective, the domain of online education provides valuable resources in the form of vast amounts of textual, behavioral and structural data that can aid pedagogical tools. At the same time, advances in the fields of Natural Language Processing and statistical machine learning provide us tools to explore massive text and interaction data. In the recent past, research has focused on leveraging this data to understand learning behavior such as adaptive language teaching (the ongoing IBM confidential "HLT Assisted Language Learning" project at the IBM TJ Watson Center about which I learnt during my intership at IBM), Educational data mining, modeling student learning [9, 10], and predicting student behavior [18] etc.

My research focuses on tapping the potential for statistical text analysis in this area, and I am excited by the possibility of its far-reaching potential benefits. I plan to use probabilistic modeling and text mining to address some of the vital problems in this domain, and use the data to infer behavioral and didactic insights. I aim to build learning systems that improve the online learning framework in four ways:

- Automatically analyzing textual content from student responses, questions and forum-posts to quantitatively measure class comprehension
- Automatically identifying if an answer to a previously unseen subjective question is relevant
- Statistically identifying non-verbal cues and behavior patterns to identify a student's interests and predicting student performance and dropping out intentions before course completion
- Adapting data from different courses to improve system-performance, and predicting performance of trained predictive models when tested on newer courses

Evaluating class comprehension

Enrollment in a typical MOOC ranges from a few hundred to upto tens of thousands of students, while the instructional staff size is much, much lower. The bloated ratio of students to staff in the online education setting immediately calls out for automated/semi-automated methods for analysis of students' submissions to assignments, forum questions and posts. While there has been some recent work on improving peer grading methods [17, 19], it is perhaps more important for the teaching staff to analyze students' understanding of the course content well before the course completion. This is even more significant for MOOC like courses provided by organizations such as Khan Academy, Peer-to-Peer University (P2PU) and Course Hero, which promise a personalized learning environment. In a traditional brick and mortar classroom, teachers interact with students and grade student responses directly, and this gives them an estimate of overall students understanding of various concepts. In MOOCs, since the interaction is virtual and the assignments are automatedly or peer-graded, the only way for an instructor to analyze students' understanding is to manually go through the discussion forums and submissions. This, in itself, is a very challenging task. For example, a recent MOOC offered by UMD had 16K students and more than 11K posts on the dicussion forum, apart from thousands of submissions for weekly quizzes and assignments. My ongoing work of understanding collections of similar documents is directed towards this problem. We are currently developing a non-parametric Bayesian model extending Hierarchical Dirichlet Processes (HDP) to automatically identify fine-grained aspects present in a set of similar documents such as students' responses to a quiz/assignment/forum question. The model identifies the major underlying aspects present in the data and labels sentences with aspects. This can possibly be helpful in evaluating comprehension in two distinct unsupervised ways:

1. It can help to automatedly evaluate a student response by comparing its aspects with those from a set of model responses.

2. It can also be used in summarizing the gist of several documents concisely in a few sentences by picking representative sentences that maximally cover important aspects in the document set, and hence provide a succinct description of overall class understanding.

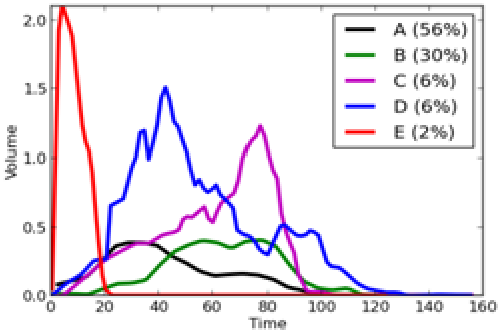

Figure 1: Posting frequency patterns of students with varying grades for Stanford machine learning MOOC

Our initial experiments on a manually annotated aspects test set have shown that our introduced model considerably improves the mean Adjusted Rand Index (a measure quantifying the coherence of aspect-clusters with human annotations) from 0.05 for HDP to 0.22. This indicates that the model can meaningfully identify aspects present in a text. This can then be analyzed by the course staff to quantify the class understanding of a course module, and to possibly modify lecture content and course progression accordingly.

Question Answering

Due to the nature of interactions in online courses, most communications are in the form of questions and answers (mostly involving subjective questions). Especially, in lieu of face-to-face communication, MOOCs have highly active question-answering forums which serve as useful platforms for seeking answers and clarifications. However, in general, only a small fraction (approximately 10%) of the students use these forums regularly. Since these forums are large and grow quickly, it is extremely difficult for a non-regular forum user to follow threads and find answers to the question they might have. Simply browsing the numerous thread titles and lexical searches

“I agree with everyone above on Surviving Disruptive Technologies being a great course! This has been such a nice and novel experience for me..” (Score - 84.6/100)
“I like this course and i have learned a lot. It is good organized and the video’s have a high value.” (Score - 71.1/100)
“I’ll admit it. I hate this class/course...” (Score - dropped)

Figure 2: Posts and scores of some students from the Disruptive Technology MOOC offered by UMD

don’t suffice as discussions on these threads are informal and often go off-topic. There is, hence, a substantial need to automatically analyse forum data and identify relevant answers to subjective questions posed by an non-regular forum user.

My recent work during an internship at IBM Watson (summer 2013) applies to these settings. We addressed the problem of predicting if a given answer snippet is relevant to a subjective question. We developed data-driven joint probabilistic models that assume that questions belong to latent types or clusters, and hence simultaneous clustering and building cluster specific relevance prediction models would be a better approach than having global cluster agnostic models [12]. Our experiments with two question answering datasets have yielded promising results in question clustering. Our joint model (F1 score=49.80) significantly outperformed the traditional system (F1 score=45.99) when the upper bound on the F1 score, obtained using extensive human annotations, was 50.87. Similar results were obtained on the other dataset. This approach is inherently suitable for online education forums where the textual data we have is primarily in form of questions and subjective responses.

Student Intentions & Improving Retention

While my ongoing work is primarily based on textual analysis of student responses, there is a lot more potential and value in using an eclectic range of other behavioral and linguistic cues which can be analyzed to provide insights into student behavior and intentions. Data about students’ usage patterns, contents and sentiments of their postings and the nature of their interactions can not only help identify sentiments such as confusion or frustration to modify teaching strategy, but assess their engagement and motivation.

This is especially relevant for MOOCS that often have notoriously low completion rates due to a range of factors such as inflated expectations from the course, steep learning curves, unexpected rigor or simply finding lectures too boring etc. Studies have shown that even though students in web-based course offerings do as well or better than in traditional courses, they are often less satisfied[3]. One of our datasets reveals that out of more than about 16,000 registrants for a recent MOOC on Disruptive Technologies offered by UMD, only 800 actually completed the course. A helpful first step towards analyzing this behaviour would be to predict student performance during the initial stages of the course. In future work, I plan to look into the problem of predicting student grades and more importantly, students course completion intentions. Identifying students who are thinking of dropping-out and a causal analysis of their intention would assist the staff in identifying problem areas and working towards its solution by providing extended explanations, extra practice questions at an academic level; or creating collaborative group assignments and personally reaching out to create a feeling of community that traditional classrooms often have.

A recent work by Kizilcec et al. [18] uses simple K-means clustering on patterns of interaction with videos and assessments to categorize students into four clusters describing their level of engagement. While such research could be helpful in understanding the demographics of the course, considering the rich nature of data accompanying MOOCs, there is often a hoard of other information that can be harnessed to better understand student synamics, and make better predictions. These include pattern of posting on the course forum, performance in the initial quizzes and assignments, behavioral patterns of re-watching videos or re-attempting quizzes, structural features from students social networking within the classroom etc.

For a simple example, Figure 1 shows the forum posting frequencies for students who got different grades in a popular MOOC offered by the Stanford University. There is a clear distinction between forum posting frequencies of students who got good grades and bad grades and modeling this simple pattern could help predict student performances. Figure 2 shows posts by students who obtained good scores and by those who dropped the course. Clearly, textual analysis of these posts can be valuable for predicting students' performance and for identifying students who intend to drop the course.

Also, Figure 3 shows a graph of students enrolled in a MOOC in Disruptive Technology offered by UMD. Each node is a student and there is an edge from node A to node B if A posted oon a thread started by B. The nodes were clustered using Clauset-Newman-Moore, a topological clustering algorithm, and were colored and sized by grades. Hence, small yellow nodes represent students who didn't perform well while bigger and darker nodes represent students who scored well. We can see that G1 on top left corner of the graph consists of mostly big blue nodes highly connected to each other while G4 towards the bottom middle consists of small yellow nodes not connected to others. This indicates that graph clusters can help in predicting final score for a student.

In oncoming work, I plan to build probabilistic machine learning models that would rank students in order of their predicted performances by incorporating structural as well as behavioral cues. The technical challenge here lies in elegantly incorporate assorted collections of features into predictive models, while retaining their intuitive meanings. I have worked on a discriminatively enhanced topic model that enables easy incorporation of hand-crafted features in a V-structure (when the predicted variable depends on multiple sources) by embedding a log-linear component. Our experiments revealed that such an enhancement improved the predictive performance of the model by providing more flexibility in the range of features incorporable and reducing the number of parameters to be learnt while retaining the generative capabilities [13]. Hybrid latent variable approaches like these could be especially valuable for the domain of online education where we not only want a model with good discriminative ability, but also interpretable explanations of the prediction that can be used to improve didactic strategies.

Adaptation

In addition to my ongoing work that focuses on text analysis and probabilistic modeling, I briefly outline a challenging problem that is extremely relevant, and calls for distinctly different techniques. This relates with the issue of adaptation, in two senses of the word. Firstly, to provide a personal educational experience tailored to a student's interests and abilities, it is necessary to dynamically adapt the course-content presented to him/her. This extends beyond enabling a flexible course schedule and letting students watch videos or take quizzes at their own speed. Due to the open nature of these courses, there is considerable variation in the background, skill set and potential of students. It would be beneficial to model the weak and strong areas for individual students and suggest additional study material for the former while skimming over topics of student's expertise. This would retain the students' interest, save their time and improve their experience and possibly outcome. While rule-based methods could be used to adapt the system response according to a student's performance and other cues gathered from his profile [15], in the longer term it might be more conducive to learn this adaptation in a data-driven way (e.g., formulating this as a reinforcement learning problem where the reward-function corresponds to student performance in course-evaluation).

Secondly, and more urgently, there is a need for domain adaptation to build better models for courses that don't have enough data, using data from other sources. This is the case, for instance, for courses with no previous offerings or those with small class-sizes. Most statistical learning algorithms assume that the training and the test data are similarly distributed. This means that models trained for a course might not directly perform well for another course. MOOCs started in 2008 with a course in 'Connectivism and Connective Knowledge' and now span a diverse spectrum of fields. Within a year, Coursera alone had offered about 325 courses, with 30% in the sciences, 28% in arts and humanities, 23% in information technology, 13% in business, and 6% in

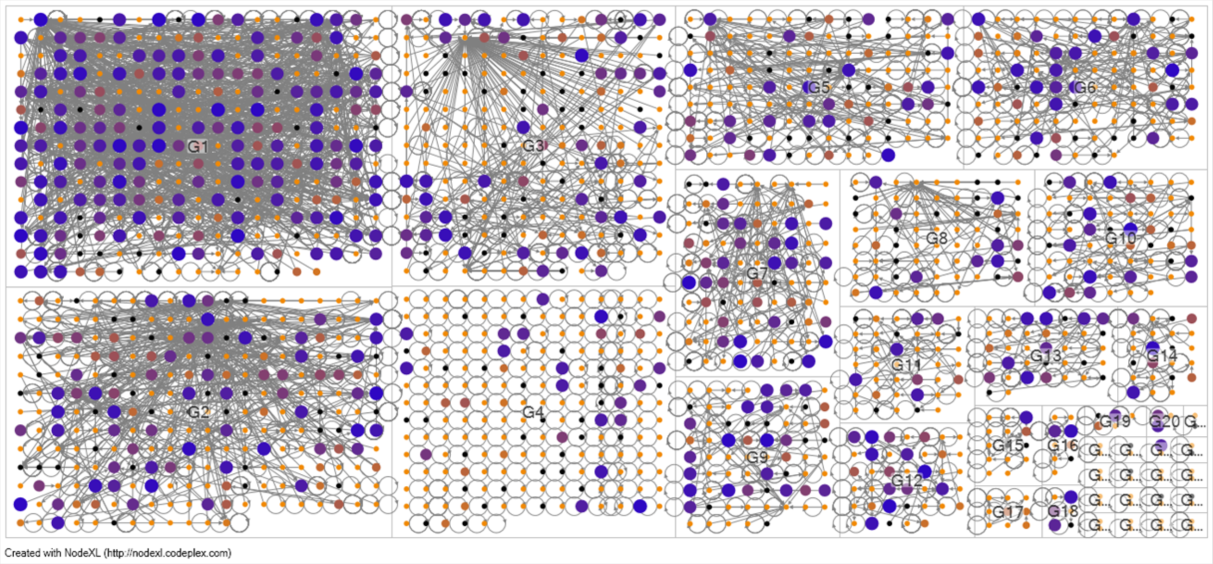

Figure 3: Social network graph for students from a MOOC in Disruptive Technology offereed by UMD

mathematics [20]. Since the nature of data for such diverse courses is expected to differ, techniques from the field of domain adaptation [16] would be very handy for solving this problem. Earlier, as a Blue Scholar at IBM Research, India, I worked on estimating variance in accuracy of text-processing systems on unlabeled datasets from other domains. In theory, the same approach could be used to estimate the expected variance when using a model trained on one course to make predictions about other courses. We had characterized datasets using meta-feature signatures, and used these to estimate a model's accuracy on a new corpora using meta-regression [14]. Our approach had yielded encouraging results across tasks such as address segmentation and document classification, and for both rule based and statistical learning based systems. In the future, I would like to extend the idea to discriminative models in the domain of online education. Specifically, I would like to estimate performances of predictive models trained using one course on new courses and analyze how they can be generalized to fit well on newer courses.

Applications to IBM

My research focus aligns closely with IBM's 'Education for Smarter Planet'[4] initiative within IBM's smarter planet theme. My work primarily aims at exploiting the massive amounts of textual and behavioral data to improve learner's experience and academic outcome. This would be directly relevant to IBM's cloud-based 'Framework for smarter education' which is geared towards monitoring student performance using predictive modeling methods. My current work could also find immediate application in the 'Data and analytics for Smarter Education'[5] solution that aims to uncover latent associations from heterogenous data sources to "predict student outcomes and recommend early interventions." Such solutions could benefit from predictive models trained on textual, structural and behavioral data. Additionally, generative components in these models can make the solutions interpretable and much easier to explain. An underlying theme in all of these solutions is to use all of a learner's information to respond according to his/her needs. My specific drive towards analyzing the textual content of students' submissions and posts for a better understanding of general class comprehension and also improving student retention fits in this setting. In addition to these, several other IBM projects whose goals relate with my research include the ‘IBM Business Analytics software for education’[6], the ‘Personalized Education Through Analytics on Learning Systems (PETALS)’[7], ‘EduPaL: Smarter Education Analytics and Delivery’[8] projects and the ‘Adaptive systeM for Personalized LEarning (AMPLE)’ projects at IBM Research, India and the ‘HLT Assisted Language Learning’ project (IBM confidential) at the IBM TJ Watson Center. These soutions provide dashboards for easy visualization, exploration and summary of data, which could directly incorporate my research.

Conclusion

MOOCs are an exciting new educational paradigm, posing several pedagogical and computational challenges, but provide vast volumes of heterogenous data. In my PhD research, I plan to analyze and address the problems of scale in online education using the massive amounts of data of unprecedented textual, behavioral and structural nature that the domain offers. I believe that learning algorithms and unsupervised methods can not only improve the domain of online education, but possibly also lead to observations that can improve our understanding of pedagogy. While I understand that the issues I highlight are only a subset of the considerable challenges in this domain; they are pertinent as well as interesting from a research perspective. With the ever-growing popularity of online education, these issues are bound to become more serious in time.

References

[1] http://nation.time.com/2012/10/18/college-is-dead-long-live-college/.

[2] http://www.nytimes.com/2012/07/20/opinion/the-trouble-with-online-education.html.

[3] http://www.westga.edu/~distance/ojdla/fall53/rivera53.html.

[4] http://www.ibm.com/smarterplanet/us/en/education_technology/ideas.

[5] http://www.ibm.com/smarterplanet/us/en/education_technology/nextsteps/solution/K492348W87462Z15.html.

[6] http://www.ibm.com/smarterplanet/us/en/education_technology/nextsteps/solution/S292500W43329H30.html.

[7] http://researcher.ibm.com/researcher/view_project.php?id=4972.

[8] Edupal. http://researcher.ibm.com/researcher/view_project.php?id=4976.

[9] M. Berland and T. Martin. Clusters and patterns of novice programmers. In *AREA*, 2011.

[10] P. Blikstein. Using learning analytics to assess students’ behavior in open-ended programming tasks. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, LAK ’11, pages 110–116, New York, NY, USA, 2011. ACM.

[11] L. Bomia, L. Beluzo, D. Demeester, K. Elander, M. Johnson, and B. Sheldon. The impact of teaching strategies on intrinsic motivation. *ERIC Clearinghouse on Elementary and Early Childhood Education*, page 294, 1997.

[12] S. Chaturvedi, V. Castelli, R. Nallapati, H. Raghwan, and R. Florian. Probabilistic models for relevance prediction in non-factoid question answering (under preparation).

[13] S. Chaturvedi, H. Daume III, and T. Moon. Discriminatively enhanced topic models. In *ICDM*, 2013.

[14] S. Chaturvedi, T. A. Faruquie, L. V. Subramaniam, and M. K. Mohania. Estimating accuracy for text classification tasks on large unlabeled data. In *CIKM*, pages 889–898, 2010.

[15] S. Chaturvedi, K. H. Prasad, T. A. Faruquie, B. S. Chawda, L. V. Subramaniam, and R. Krishnapuram. Automating pattern discovery for rule based data standardization systems. *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, 0:1231–1241, 2013.

[16] H. Daumé III and D. Marcu. Domain adaptation for statistical classifiers. *J. Artif. Intell. Res. (JAIR)*, 26:101–126, 2006.

[17] I. Goldin. Accounting for peer reviewer bias with bayesian models. In *In Proceedings of the Workshop on Intelligent Support for Learning Groups at the 11th International Conference on Intelligent Tutoring Systems*, 2012.

[18] R. F. Kizilcec, C. Piech, and E. Schneider. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, LAK '13, pages 170–179, New York, NY, USA, 2013. ACM.

[19] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller. Tuned Models of Peer Assessment in MOOCs, July 2013.

[20] M. Waldrop. Massive open online courses, aka moocs, transform higher education and science. In *Nature Magazine*, 2013.