Visual Analytics for MOOC Data

Huamin Qu and Qing Chen
Hong Kong University of Science and Technology

In the last several years, massive open online courses (MOOCs) have emerged and attracted a remarkable amount of public attention. With the rise of MOOC e-learning platforms, learners from all over the world can now enroll in more than 1,000 courses, and the number of MOOC registrants has reached 10 million. Based on our observations of this emerging movement, we believe that MOOCs provide an opportunity for visualization researchers and that visual analytics systems for MOOCs can benefit a range of end users such as course instructors, education researchers, students, university administrators, and MOOC providers.

Major MOOC platforms such as Coursera and edX can provide raw data to course instructors and their collaboration partners. In addition to the basic information in learner profiles, MOOC platforms also record web log data such as video viewing histories, clickstreams of course videos, and activities in course forums. The performance of the students as well as their genders, ages, and countries are also available. It is the first time in history that such comprehensive data related to learning behavior has become available for analysis. Compared with traditional education records, MOOC data has much finer granularity (more activities from many more students) and contains new pieces of information (such as "seek" click events in clickstreams).

However, challenges arise when analyzing MOOC data. First, the data is large, complex, and heterogeneous. It is common for one course to have more than 10,000 registrants, generating millions of clickstream events. The data contain structured, unstructured (such as text in course forums), spatial, and temporal information. In some cases, the data might also be sparse or noisy. For example, because instructors and teaching assistants (TAs) have no way to grade tens of thousands of student quizzes, peer grading is widely used in MOOCs. The data from peer grading is both sparse (for example, each student only grades a few questions) and noisy (for instance, the quality of the grading may not be reliable). Thus, it is challenging to extract meaningful information from this kind of data.

Second, the end users of the analytics systems such as course instructors, education researchers, and students usually have little or no knowledge of data mining techniques. Therefore, it is critically important to offer them an easy-to-use analytic system with clear visual aids to lower the learning curve and help them analyze the data more intuitively and efficiently. Visualization techniques have to be used to provide an intuitive interface, but those visualizations have to be well chosen such that the end users can easily grasp them without much training.

Third, the different MOOC user group needs are diverse. For example, course instructors and MOOC video providers want to know if the lecture videos are engaging and which parts of each video excite viewers and which parts are skipped. Course instructors may also want to monitor the performance and sentiments of students and identify the challenges facing different student groups. Education researchers, on the other hand, want to understand the reasons behind the high drop-out rates in MOOCs and evaluate the effectiveness of different assessment schemes (such as peer grading). MOOC platforms want to know whether the forum provides an effective way for students to communicate with instructors, TAs, and other students. Students may want to review their use of course materials and monitor their performances compared with other students taking the same course. Lastly, university administrators want to know whether the data from MOOCs can provide a new way to evaluate instructors.

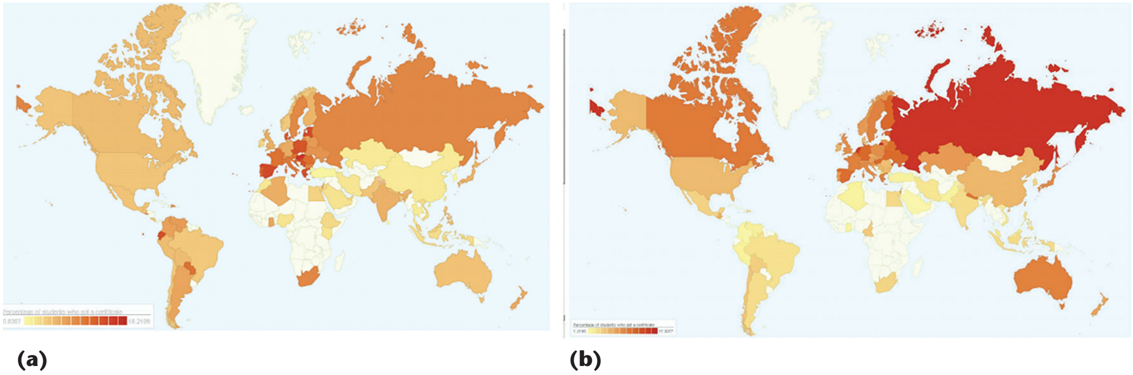


Figure 1. MoocViz heatmap visualizations. The ratios for the number of students who gained certificates versus the number who registered for two courses: (a) 6.002x course offered by the Massachusetts Institute of Technology via edX, and (b) Crypto 1 course offered by Stanford University via Coursera. (Image courtesy of Franck Dernoncourt and his colleagues.[1])

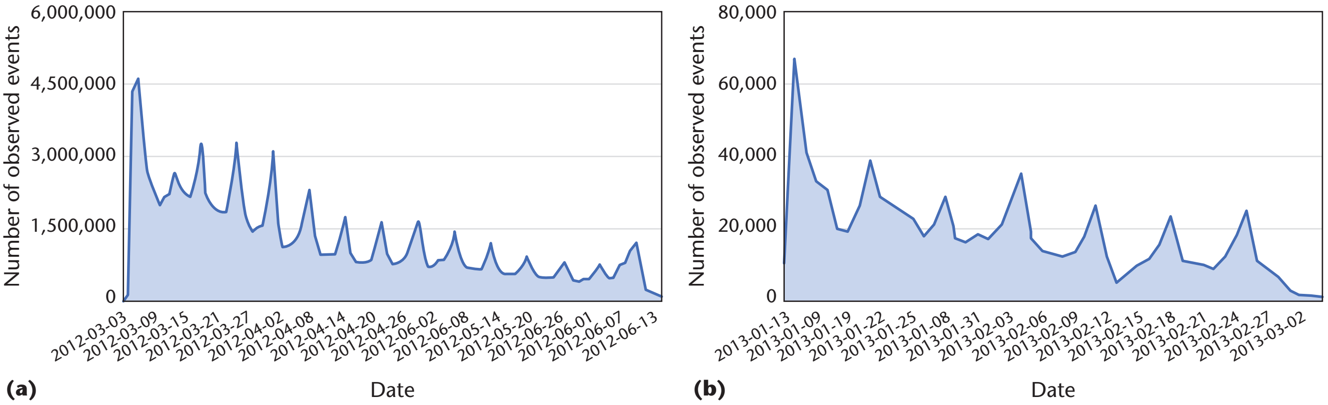


Figure 2. MoocViz line chart visualizations. The number of observing events by day for the two courses: (a) MIT's 6.002x course via edX, and (b) Stanford's Crypto 1 course via Coursera. (Image courtesy of Franck Dernoncourt and his colleagues.[1])

To address these challenges, we need to develop both advanced data mining methods to reveal patterns from MOOC data and visualization techniques to convey the analytical results to end users and allow them to freely explore the data by themselves.

Current Progress

Although MOOC data have only become available recently, some visualization tools and visual analytic systems have already been developed. These tools or systems let users analyze data with various visual aids and rich interactions. MoocViz[1] provides a cross-course, cross-platform analytics framework to accommodate the diverse needs of five types of users: course instructors, education researchers, arms-length observers, the technology-savvy crowd, and MOOC providers. The framework utilizes a data model that can adapt to different MOOC platforms, and mainstream analytical tools such as Matlab can also be easily integrated. However, general end users such as course instructors usually have limited knowledge of programming, so they may find it difficult to use the provided analytical tools to explore the data and analyzing learning behaviors.

MoocViz incorporates several basic visualizations such as heatmaps and line charts to present aggregate statistics such as the ratio of the certificate achievers to the number of registrants. Figures 1a and 1b show the certificate achiever ratio by country for two MOOC courses offered by the Massachusetts Institute of Technology and Stanford University, respectively. From the heatmaps, we can see that the three countries with the highest certificate achiever ratio for the MIT 6.002x course are Hungary, Spain, and Latvia, whereas the three countries with the highest ratio for the Stanford Crypto 1 course are Russia, the Netherlands, and Germany. Figure 2 shows the change in the number of observing events along the course timeline. These visualizations are useful but inadequate because end users also want to know the reasons behind those different behaviors.

Carleton Coffrin and his colleagues investigated how students engage with MOOCs using

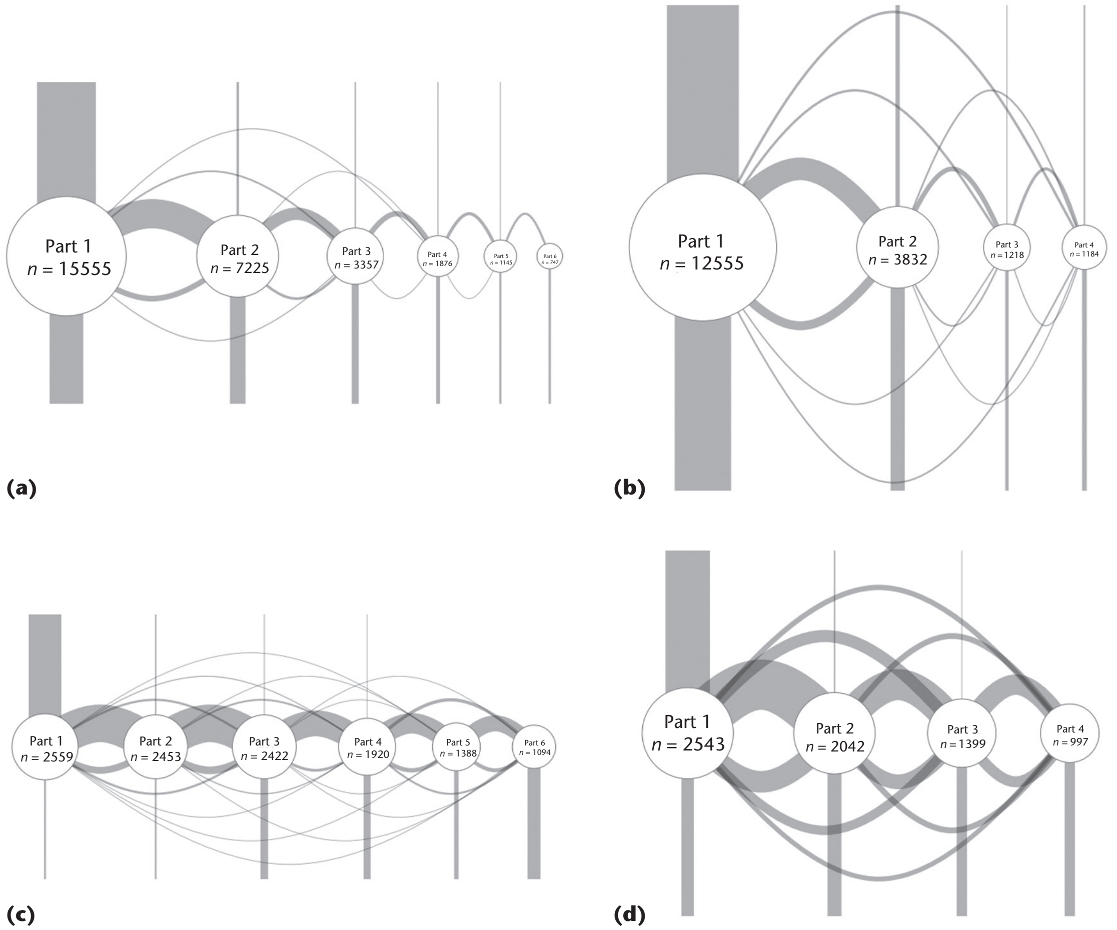


Figure 3. Student video viewing transitions by two subgroups: nonqualified and qualified students. The four state transition diagrams show data from two different courses developed at the University of Melbourne: (a) Principles of Macroenomics course (nonqualified), (b) Discrete Optimization course (nonqualified), (c) Principles of Macroenomics (qualified), and (d) Discrete Optimization course (qualified). (Image courtesy of Carleton Coffrin and his colleagues.[2])

two MOOCs developed at the University of Melbourne: "Principles of Macroeconomics" and "Discrete Optimization." The data they used included student participation (video views and assignment submissions), assessment performance (marks), and online interactions (every interaction a student has with the MOOC platform). Apart from a series of basic visual presentations, such as bar and line charts, they used state transition diagrams to depict students' interactions and to look for engagement patterns and their relation with student performance. Their findings demonstrate the power of various visualization techniques in presenting learning analytics results.

Despite the differences between the two University of Melbourne courses in terms of curriculum and assignment design, there are still some similar patterns such as the cumulative distributions of student performance and weekly student participation by student subgroups. The state transition diagrams in Figure 3 demonstrate the benefit of visualizing learning analytics results. The figure shows that different subgroups of students have different transition patterns for MOOC videos. This work also revealed that the students' activities and performances at the beginning of a course significantly affect their final outcomes. However, the detailed information was lost in the overall statistical analysis, preventing users from drawing more insightful conclusions.

Other researchers have studied how video production affects student engagement and the relation of in-video dropouts and interaction peaks in the lecture videos.[3,4] The data in these studies[3,4] include 6.9 million video watching sessions from four courses at edX. Based on their findings, the researchers made several recommendations to help instructors and MOOC video producers produce better lecture videos. The analysis in Figure 4 reveals that students who rewatch videos are more likely to drop out. This might indicate that students who rewatch material have a more specific purpose,

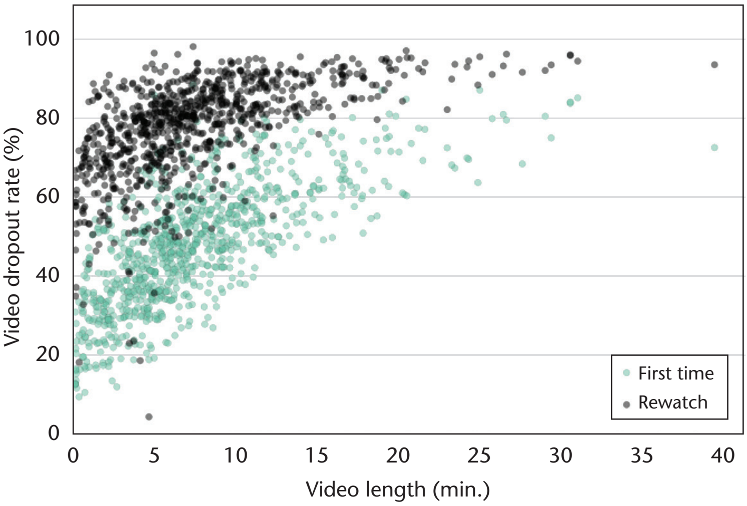


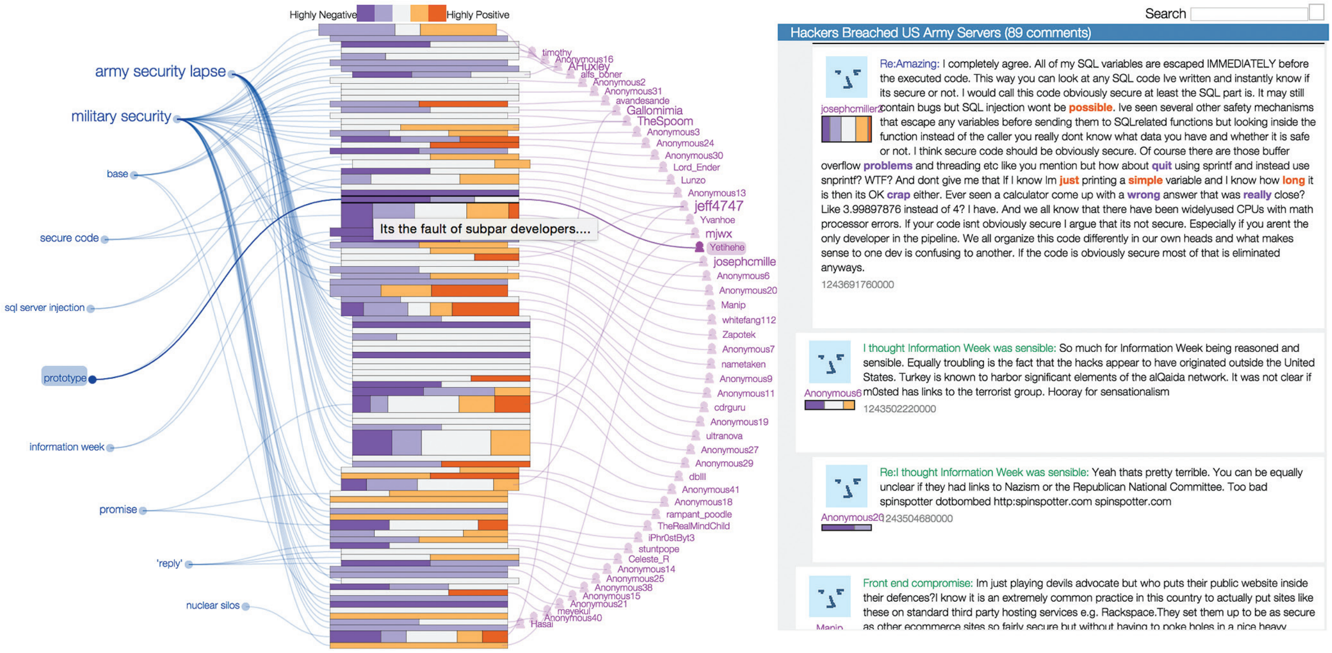
Figure 4. Student dropout rates. Students who rewatched videos (78.6 percent, with a standard deviation of 54.0) tended to drop courses at a much higher rate than single watchers (48.6 percent, with a standard deviation of 34.0). (Image Courtesy of Juho Kim and his colleagues.[4])

so they only watch select videos. The authors also obtained several other interesting findings: most students watched online course videos for no more than 6 minutes (meaning the videos could only hold their attention for this amount of time), and courses with longer videos usually had a higher dropout rate. Some proposed course video design guidelines thus included creating shorter videos (such as by making video segments shorter than 6 minutes) and avoiding abrupt visual transitions.

In addition to user profile data and student interactions with course videos, the data from forums, where students can write posts and communicate with each other, is also valuable for e-learning behavior analysis. Unlike traditional face-to-face teaching, MOOC instructors do not get direct feedback from students. Therefore, forums play a key role in communication between instructors and students. Recently, a visual analytic system, ConVis,[5] has been proposed to visualize conversational data in blogs (see Figure 5). Text structures, authors, and topics can be shown together in a relatively small space, providing a compact summary of a conversation. Interactions such as brushing and highlighting are implemented to help users explore the system more efficiently. However, it is difficult for ConVis to visualize the relatively large-scale forum data from MOOCs. In addition, cross-thread relationships cannot be shown in ConVis. More efforts are needed to develop similar visualizations for MOOC forum data.

In earlier work, we developed VisMOOC,[6] a comprehensive visual analytic system that helps MOOC instructors and education researchers analyze the clickstream data and forum data from MOOC platforms. The system was developed together with five MOOC instructors and one education analyst based on data from three MOOCs offered by the Hong Kong University of Science and Technology on both Coursera and edX.

Figure 6 shows the VisMOOC interface. It contains a list view for selecting a MOOC course and the videos of the chosen course on the left, a



Hackers Breached US Army Servers (89 comments)

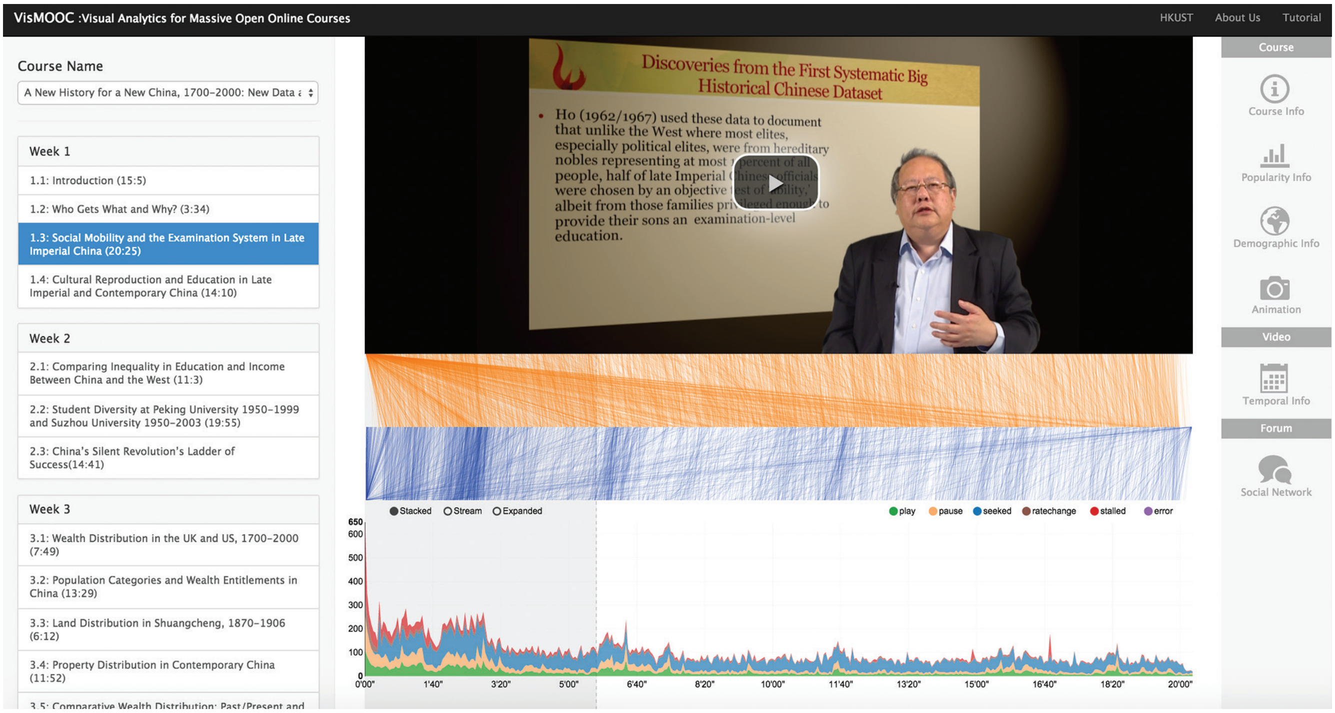
josephcmiller2 — Re:Amazing: I completely agree. All of my SQL variables are escaped IMMEDIATELY before the executed code. This way you can look at any SQL code Ive written and instantly know if its secure or not. I would call this code obviously secure at least the SQL part is. It may still contain bugs but SQL injection wont be **possible**. Ive seen several other safety mechanisms that escape any variables before sending them to SQLrelated functions but looking inside the function instead of the caller you really dont know what data you have and whether it is safe or not. I think secure code should be obviously secure. Of course there are those buffer overflow **problems** and threading etc like you mention but how about **quit** using sprintf and instead use snprintf? WTF? And dont give me that If I know Im **just** printing a **simple** variable and I know how **long** it is then its OK **crap** either. Ever seen a calculator come up with a **wrong** answer that was **really** close? Like 3.99897876 instead of 4? I have. And we all know that there have been widelyused CPUs with math processor errors. If your code isnt obviously secure I argue that its not secure. Especially if you arent the only developer in the pipeline. We all organize this code differently in our own heads and what makes sense to one dev is confusing to another. If the code is obviously secure most of that is eliminated anyways.
1243691760000

Anonymous6 — I thought Information Week was sensible: So much for Information Week being reasoned and sensible. Equally troubling is the fact that the hacks appear to have originated outside the United States. Turkey is known to harbor significant elements of the alQaida network. It was not clear if m0sted has links to the terrorist group. Hooray for sensationalism
1243502220000

Anonymous20 — Re:I thought Information Week was sensible: Yeah thats pretty terrible. You can be equally unclear if they had links to Nazism or the Republican National Committee. Too bad spinspotter dotbombed http:spinspotter.com spinspotter.com
1243504680000

Manip — Front end compromise: Im just playing devils advocate but who puts their public website inside their defences?I know it is an extremely common practice in this country to actually put sites like these on standard third party hosting services e.g. Rackspace.They set them up to be as secure as other ecommerce sites so fairly secure but without having to poke holes in a nice heavy

Figure 5. ConVis. This visual analytics system developed to visualize conversational data in blogs shows the text structures, authors, and topics in an integrated display. (Image Courtesy of Enamul Hoque and Giuseppe Carenini.[5])


VisMOOC :Visual Analytics for Massive Open Online Courses

Course Name

A New History for a New China, 1700–2000: New Data a

Week 1

1.1: Introduction (15:5)

1.2: Who Gets What and Why? (3:34)

1.3: Social Mobility and the Examination System in Late Imperial China (20:25)

1.4: Cultural Reproduction and Education in Late Imperial and Contemporary China (14:10)

Week 2

2.1: Comparing Inequality in Education and Income Between China and the West (11:3)

2.2: Student Diversity at Peking University 1950–1999 and Suzhou University 1950–2003 (19:55)

2.3: China's Silent Revolution's Ladder of Success(14:41)

Week 3

3.1: Wealth Distribution in the UK and US, 1700–2000 (7:49)

3.2: Population Categories and Wealth Entitlements in China (13:29)

3.3: Land Distribution in Shuangcheng, 1870–1906 (6:12)

3.4: Property Distribution in Contemporary China (11:52)

3.5: Comparative Wealth Distribution: Past/Present and

Discoveries from the First Systematic Big Historical Chinese Dataset

- Ho (1962/1967) used these data to document that unlike the West where most elites, especially political elites, were from hereditary nobles representing at most 1 percent of all people, half of late Imperial Chinese officials were chosen by an objective test of ability, albeit from those families privileged enough to provide their sons an examination-level education.

Figure 6. VisMOOC system interface.[6] **The list view is on the left, the content-based view is in the middle, and the dashboard is on the right.**

content-based view in the middle, and a dashboard on the right.

In the content-based view, users can analyze the clickstream data together with the corresponding video content. An event graph and a seek diagram present different types of clickstreams along the video timeline. The event graph displays the distribution of click events to help users study how e-learners interact with a certain video. Colors are used to encode six different types of click events, and the height shows the number of events. Among the six types of click events recorded by the Coursera platform, instructors found the seek events most interesting because they can reveal which parts students rewatched and which parts they skipped. A forward seek from an earlier time point to a later time point in the video shows that a student skips a certain part of the video, while a backward seek from a later time point to an earlier time point indicates that a student rewatched a segment. Therefore, we designed a seek diagram to present the starting and ending points of forward and backward seeks, respectively. A line between the two axes connects the starting and ending positions together for each seek event. A cluster of dense lines in a backward seek diagram indicates that many students reviewed that part, which is worth further investigation. Experts can then click on the corresponding video segments to explore the reasons behind those reviewed parts.

The dashboard view allows users to analyze the data from different perspectives and at multiple levels. At the course level, the system provides the course information, the popularity of each video, and the demographic distribution of the students. Users can click on a particular country to select students from that country for further analysis and then only the selected students will be shown in the seek diagram and event graphs in the content-based view. Users can also play an animation to show the click actions that occurred for all videos during the whole course period. At the video level, the system provides a temporal information view that displays how much the popularity of a selected video varied from day to day.

When the instructors explored the demographic information in the dashboard view, they wondered whether students from different countries had different clicking behaviors. Thus, they filtered the event graphs by selecting different countries. Figure 7 shows the clickstream distributions of students from the US and China. By comparing the two event graphs, the instructors found that the percentage of seek events for the US is much larger than that of China. The experts then tried to explain this phenomenon. Some instructors observed that, based on their own experiences from face-to-face teaching, Chinese students tend to take notes. Thus, the Chinese students might prefer to pause the video to take notes, while students from the US skip more often.

To explore the MOOC data from a forum perspective, VisMOOC also provides social network analysis and sentiment analysis for the forum data in the dashboard view. Figure 8 shows the forum

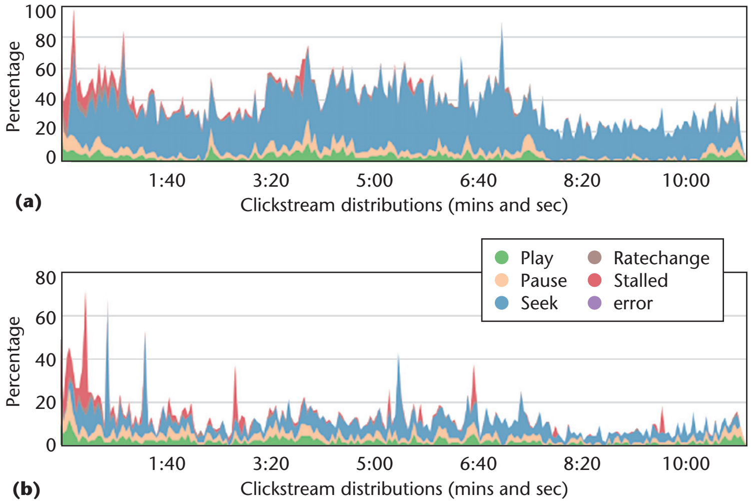


Figure 7. Clickstream differences between countries. Event graphs for students from (a) the US and (b) China.

social network according to students' posting interactions. Each dot represents a student, and an edge connects two dots if one student answers another student's post. The size of the dot indicates the student's activity level, and the color shows the student's grade. By hovering over a certain dot, users can view the student's personal network in the forum. A few students actively participated in the forum but didn't receive a grade. Some of these people were socializers, whose goals were not necessarily to learn from the course, while others turned out to be the course TAs using the system to answer students' questions.

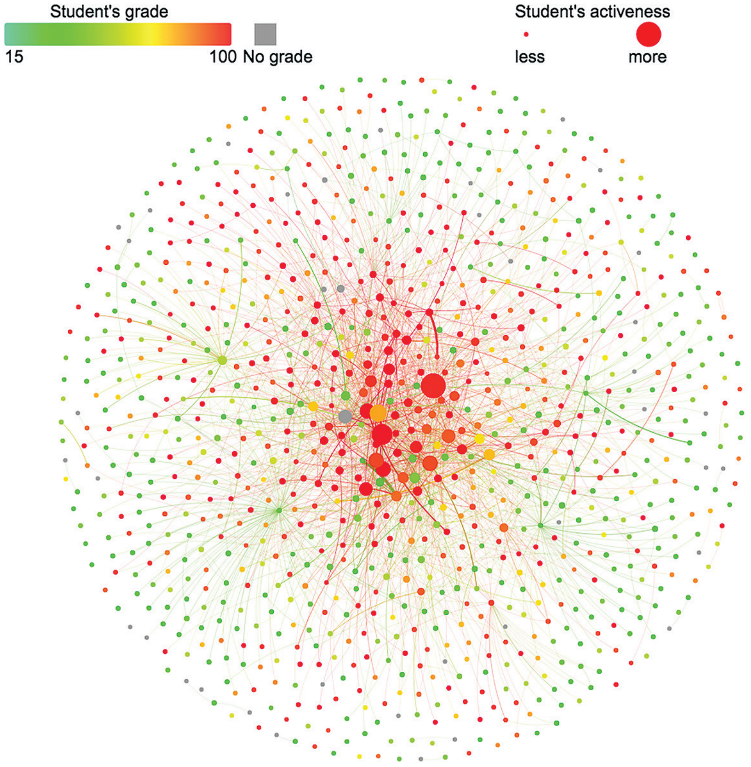


Figure 8. VisMOOC forum social network analysis. The size of the dot shows the student's activity level, and the color shows the student's grade. Hovering over a dot highlights that student's personal network.

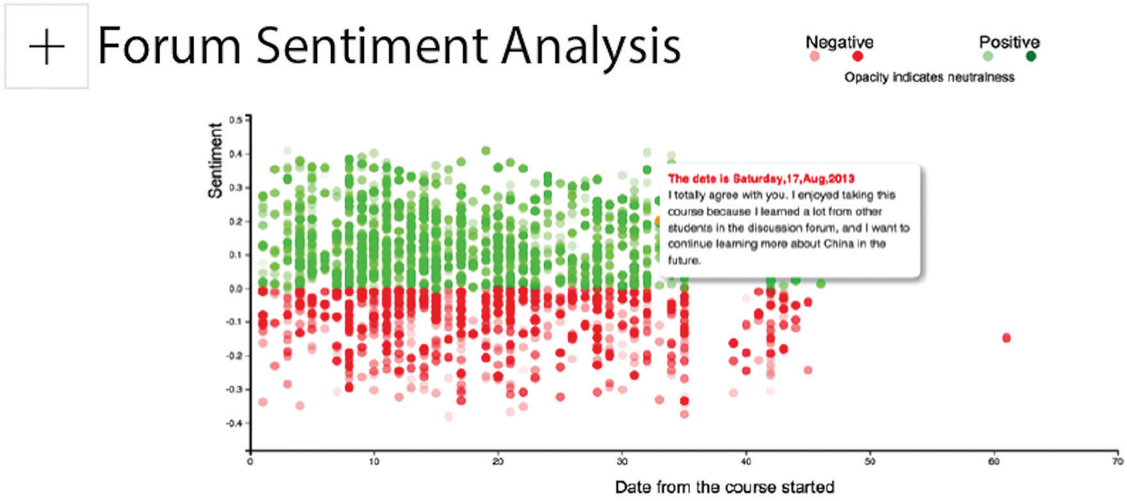
Figure 9 shows the sentiment analysis for the forum data using the Python NLTK Text Classification (http://text-processing.com/demo/sentiment/), where green dots indicate positive posts and red dots are negative posts. The y-axis value shows the extent to which the posts are positive or negative. By hovering on a specific dot, users can view each post's detailed information. With this tool, the course instructors can easily monitor the sentiments of students.

Call for Action

The various examples we have explored here illustrate some typical visual analytics systems developed for MOOC data. From these examples, we can see that the availability of MOOC data provides a great opportunity to understand learning behaviors, and the visualization techniques can play an important role in analyzing such data. The insights gained into online learning behaviors by using these systems can help instructors design better course content.

Yet, visual analytics of MOOC data is still in an early stage and many challenges remain. First, because of data privacy issues, the raw data from MOOC platforms are only available to MOOC instructors and other authorized people. Many visualization researchers cannot access the data. Also, the raw datasets from the various MOOC platforms contain different kinds of information and come in different formats. Recently, XuetangX, a Chinese MOOC learning platform initiated by Tsinghua University, released some MOOC data for the KDD Cup 2015 competition, which asks participants to predict whether a user will drop a course within the next 10 days based on his or her prior activities. (See https://kddcup2015.com/information.html for more details.) We hope other MOOC platforms and institutions will follow this example and release more MOOC data in an anonymous and aggregated format to the research community.

Second, support for on-the-fly analysis of streaming web log data is still lacking. Currently, instructors can only access the complete data after a course is finished. Any analysis may help them improve course materials for future courses, but it would be much more desirable if instructors could do the analysis during the course period and use the analytical results to immediately improve the course content. MOOC platforms usually do not allow face-to-face interactions between teachers



The date is Saturday,17,Aug,2013

I totaly agree with you. I enjoyed taking this course because I learned a lot from other students in the discussion forum, and I want to continue learning more about China in the future.

Figure 9. VisMOOC forum sentiment analysis. The color of the dot indicates whether the post is positive (green) or negative (red), and the location of a dot indicates the extent to which a post is positive or negative.

and students. Thus, timely feedback from analytical systems is especially important to help instructors assess the effectiveness of teaching and learning and make adjustments accordingly.

Third, to achieve the goal of personalized learning, more accurate user group analysis is necessary. Different user groups might have different learning behaviors. If visual analytics can help reveal these learning behaviors, course instructors can design tailored course materials for different groups. However, this task is challenging because many different criteria can be used to classify student groups and the joint analysis of different data sources such as clickstream data and forum data is often needed to find suitable criteria.

Lastly, low retention rates are a common issue across all the MOOC platforms, and more research is needed to improve these rates via predictive analytics. Such analytics can help instructors determine which groups of students will likely drop a course so they can then take immediate action to retain those students. MOOC platforms can be redesigned based on the analytics results to offer timely feedback mechanisms, engaging forums, customized learning materials for different learning groups, and effective assessment schemes.

MOOCs offer a new education approach for both instructors and learners, but they are still at an early stage. With the improvement of MOOC platforms, we believe MOOCs will revolutionize education and provide quality, affordable education for the masses. Visual analytics can definitely play a major role during this process.

References

1. F. Dernoncourt et al., "MoocViz: A Large Scale, Open Access, Collaborative, Data Analytics Platform for MOOCs," *Proc. NIPS Workshop on Data-Driven Education*, 2013.
2. C. Coffrin et al., "Visualizing Patterns of Student Engagement and Performance in MOOCs," *Proc. 4th ACM Int'l Conf. Learning Analytics And Knowledge*, 2014, pp. 83–92.
3. P.J. Guo, J. Kim, and R. Rubin, "How Video Production Affects Student Engagement: An Empirical Study of MOOC Videos," *Proc. 1st ACM Conf. Learning at Scale*, 2014, pp. 41–50.
4. J. Kim et al., "Understanding In-Video Dropouts and Interaction Peaks in Online Lecture Videos," *Proc. 1st ACM Conf. Learning at Scale*, 2014, pp. 51–60.
5. E. Hoque and G. Carenini, "ConVis: A Visual Text Analytic System for Exploring Blog Conversations," *Computer Graphics Forum*, vol. 33, no. 3, 2014, pp. 221–230.
6. C. Shi et al., "VisMOOC: Visualizing Video Clickstream Data from Massive Open Online Courses," *Proc. IEEE Pacific Visualization Conf.*, 2015, pp. 159–166.

Huamin Qu *is a professor in the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology. His research interests include visualization and computer graphics, with focuses on urban informatics, social network analysis, e-learning, and text visualization. Contact him at huamin@cse.ust.hk.*

Qing Chen *is a PhD student in the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology. Her research interests include e-learning and human-computer interaction. Contact her at qchenah@cse.ust.hk.*

Contact department editor Theresa-Marie Rhyne at theresamarierhyne@gmail.com.