Exploring Online Learners’ Interactive Dynamics by Visually Analyzing Their Time-anchored Comments

Abstract

MOOCs are increasingly prevalent as an online educational resource open to everyone and have attracted hundreds of thousands learners enrolling these online courses. At such scale, there’s potentially rich information of learners’ behaviors embedded in the interactions between learners and videos that may help instructors and content producers adjust the instructions and refine the online courses. However, the lack of tools to visualize information from interactive data, including messages left to the videos at particular timestamps as well as the temporal variations of learners’ online participation and perceived experience, has prevented people from gaining more insights from video-watching logs. In this paper, we focus on extracting and visualizing useful information from time-anchored comments that learners left to specific time points of the videos when watching them. Timestamps as a kind of metadata of messages can be useful to recover the interactive dynamics of learners occurring around the videos. Therefore, we present a visualization system to analyze and categorize time-anchored comments based on topics and content types. Our two system integrates visualization methods of temporal text data, namely ToPIN and ThemeRiver, which could help people understand the quality and quantity of online learners’ feedback and their states of learning. To evaluate the proposed system, we visualized time-anchored commenting data from two online course videos, and conducted two user studies participated by course instructors and third-party educational evaluators. The results validated the usefulness of the approach and showed how the quantitative and qualitative visualizations can be used to gain interesting insights around learners’ online learning behaviors.

1. Introduction

Online learning platforms such as Massive Open Online Courses (MOOCs) are now widely used, and attract hundreds of thousands of learners to enroll. With the most interactions through the software platform, a rich source of information is desirable for instructors and online course producers who need to maintain and improve courses. Though it has been recommended that current online course platforms add monitoring/analysis tools including metrics tabs, Khan Academy’s Coach monitoring system, and student demographics, most such platforms still provide only limited sets of summarizations of basic, non-content features for the purposes of monitoring and analysis [SFCQ14, SMHF14]. Nevertheless, it has been reported that content-oriented qualitative information - for example, what students talk about in discussion forums is very important to instructors, as shallow qualitative features may not reveal learners’ real intentions or learning states [FZCQ17, SMHF14].

Most of the online learning platforms, especially MOOCs, apply videos as the main course material. While watching videos, learners acquire knowledge and organize them into new conception. It’s helpful for instructors to get information of how student absorb and think and how their attention and emotion vary with video contents. Guo et al. [GKR14] measured the length of video-watching time as learners’ engagements, and Kim et al. [KGS*14] investigate click-level interactions, playing, pausing, replaying, and quitting, to understand the in-video dropout rate and interaction peaks in online course videos. However, the interaction data used in these previous works mainly provide quantitative instead of qualitative information of learners’ feedbacks.

Time-anchored commenting, in which learners’ brief written observations about a video are sorted according to its time-code, has recently been applied to online educational videos [LLC*15]. Comments left by previous learners at specific time points in a video are displayed to new learners when they watch the same video and reach those time-points. As such, analyzing time-anchored commenting data could recover the interactive dynamics of learners, for example, the evolution of topics learners discussed, the temporal variation of learners’ emotion and attention and the specific events such as questions and complaints they mentioned. Therefore, to get the complete feedbacks, we need to analyze both quantitative and qualitative information of time-anchored comments.

We identified our design principles based on previous studies that investigated online learning behaviors. These principles, set forth in full in the next section, dictate that a text-analysis system be capable of categorizing students’ comments in such a way that multiple facets of each comment can be revealed. Our proposed visualization method, ToPIN, offers a qualitative view of the detailed contents of comments, such as the type of individual comments as well as topic clustering and evolution. We also utilize The-
meRiver [HHWN02], which provides an alternative quantitative view focusing on the variations in the pool of time-anchored comments. The integration of ToPIN and ThemeRiver supports global-to-local, evolitional, and multi-facet analyses that enable the analysts to visually explore the interactions among learners as well as between learners and the video lecture.

To evaluate our system as an application in the context of online education, we designed two evaluative studies, one for course instructors and the other for third-party educational evaluators. We selected these two groups as being the key stakeholders most in need of such tools in practice. In the first study, manually classified comments were fed into our visualization tool and two instructors who had produced the content of the educational videos were asked to use our system to analyze learners’ responses to their lectures. Both instructors reported very positive experiences with using our visualization.

In the second study, the comments were automatically classified by trained machine learning models in order to scale up our analysis and visualization approach. The participants in the second study included eight educational evaluators who were asked to judge online learners’ quality of learning. This study’s purpose was to verify whether a third-party researcher could easily use our system to observe learners’ learning and interactions in the context of an online course. The results indicate that our system can effectively improve researchers’ performance in analyzing learners’ responses and gaining insights into their learning. These observations can also help them to evaluate the quality of courses or provide suggestions for improving courses.

This paper is intended to make three contributions: (1) To propose a solution for visualizing both qualitative and quantitative information about online learners’ comments and demonstrating how the visualization system can be integrated into large-scale online educational videos for analyzing interactive dynamics among learners and between learners and video lecture; (2) To evaluate how visual analytics of time-anchored comments can potentially help course instructors and educational evaluators gain insights into students’ learning status; and (3) To provide design implications for comment visualization in the context of online educational videos.

2. Design principles
Stephens-Martinez et al. [SMHF14] interviewed 92 MOOC instructors about which information sources they value, and found that instructors were particularly interested in knowing the appropriateness of their classes’ difficulty levels and about learner’s engagement. Based on the requirements of instructors and educational evaluators discussed above, and the properties of time-anchored comments, we considered that the following design principles were critical.

(1) Present the temporal relation between time-anchored comments and video time. Many time-anchored comments are temporally correlated with the events in the video, and understanding the temporal patterns of learners’ behaviors when they watch an educational video is desirable for educational evaluators [KGC'14, KGS'14, SFCQ14]. The temporal variation of both quantity (numbers of comments) and quality (contents of comments) of comments gives the implicit information to know learners’ reflection on the contents in videos.

(2) Provide an overview of complex topics and present the topic-based information. It has been found that instructors wish to ascertain learners’ opinions on particular content so that they can improve their course materials [SMHF14]. Topics extracted from the learners’ comments can indicate learners’ preferences regarding the key content of videos, and will therefore offer instructors more focused insights. Besides, presenting the topic-based information - the attributes of comments (e.g., numbers, emotion, content type) in each topic - show the range and the depth it influence learners.

(3) Extract multi-faceted attributes from the content of comments. Comments are the major channel by which learners express themselves in online education. Understanding the attributes of these comments facilitates the gathering of qualitative information about students’ learning experiences by both instructors and analysts. Our visualization design should encode and exhibit the relationships and interactions between the multi-faceted content of comments and course content in an intuitive and organized way.

(4) Connect with the lecture video and comments. To allow the users to perform more detailed exploration and analysis when required, the visualization interface should integrate the visualization and the actual content of lecture videos and time-anchored comments so that the users can link visualization results to original data easily. The design to connecting visualization result and original content in a simple way is very important.

3. Related work
Our visualization system combines features of topic and conversation visualizations, and also relates to prior work on online video commenting and course visualizations. We review the relevant prior research in these areas.

Topic Visualization. Some previous approaches display topics as word clouds or text-based visual descriptions related to the frequency or timestamp of topics [LRKC10, KJL’12, WLY’14]. There are also methods that focus on visualizing the trends of topics: ThemeRiver [HHWN02] and Textflow [CLT’11] use a flowing-river metaphor to analyze thematic variation and the merging and splitting of relationships between topics. EvoRiver [SLW’14] also presents the evolution of each topic to show the cooperation and competition interactions among a set of concurrent topics on social media. TopicFlow [LMSP13] aligns similar topics across time slices and visualizes the emergence, convergence, and divergence of topics. All of these approaches are useful for understanding the variation of topics in documents or text databases. However, none of them can be directly applied to the evolution of topics in time-anchored comments, because these comments are usually short, thematically diffuse, and in some cases incoherent, as compared to the types of data that these systems were designed to deal with.

Conversation visualization focuses on identifying the temporal or structural relationship between conversations, and/or using text analysis to explore the insights generated by conversations. For example, Conversation Thumbnails [WM03] visualize...
the thread structure of conversations using tree-visualization techniques. Opinion Space [FBRG10] visualizes differences in users’ opinions by projecting them onto a two-dimensional map. ConToVi [EAGA*16] uses the ring-plot and radial visualization to show the thematic and social interaction information for multi-party conversation. ConVis [HC14] displays conversations on blogs with a thread overview that shows structure and conversational sentiments, and a facet overview covering topics and authors. An obvious difference between our method and ConVis is that we analyze and visualize the multiple facets of comments from a temporal rather than a structural perspective.

Online-video comment visualization. PolemicTweet [HIF13] provides an overview of the tags (e.g., agreement, disagreement, question, etc.) and activity spikes in audiences’ annotation of online videos. Conference Monitor (CM) [SRBS12] is a real-time, web-based tweet visualization dashboard, which monitors back-channel conversation during academic conferences. As PolemicTweet and CM were not developed for use with online educational videos, they cannot provide deeper insights related to learning experiences, such as the emotional valence and course-relevance of comments. Applying the metaphors of tree, VideoForest [SSCM16] presents the relationship between the video and comments by clustering video keyframes and grouping the comments by the codes of similarity and sentiment through the timeline. As our goal is gaining insights from learners’ feedback, our method focuses on visually analyzing the multi facets of learners’ comments and clustered topics in more details, including emotion, course-relevance, category, etc.

Online course visualization. CourseVis [MD07] provides a graphical rendering of multi-dimensional tracking data obtained from students’ log data on course websites. It helps distance-learning instructors better understand the social, behavioral, and cognitive aspects of their students’ course experiences. However, CourseVis handles neither online videos nor time-anchored comments. Because of the imprortance of forum data, iForum [FZCQ17] analyzed the activities and relation among learners to understand the participation, interaction and feedbacks from learners. Recent research has shown that students taking online courses spend the majority of their learning time watching lecture videos [SBC*14, BPD*13], leading some researchers to study learners’ behaviors when doing so: in particular, via click-stream analysis of when learners play, pause, and resume viewing lecture videos [GKR14, KGS*14, SFCQ14, SFCQ15, QCL*15, CCZ*16]. LectureScape [KGC*14] was developed to visualize and summarize lecture-video-viewers’ interaction data on a click-stream basis. It is unclear, however, whether the click stream accurately represents learners’ intentions, and the complexity and diversity of learners’ responses to course videos should also be analyzed using various other approaches.

4. Comment analysis

For design principle 3, in analyzing the multi-faceted aspects of student comments, we develop several analysis components by adopting text processing and machine learning techniques.

Emotion analysis. Instructors and educational evaluators are particularly interested in knowing if learners are satisfied with their courses, course materials, and instructors’ teaching styles. As learners often express their feelings about these aspects of their learning experiences through time-anchored commenting, our system analyzes their comments’ emotional content and categorizes it into three types - positive, neutral, and negative - using a naïve Bayes classifier.

Course-relevance analysis. Learners’ comments tend to be influenced by course content. However, due to the characteristics of time-anchored commenting, learners may also chat about topics that might not be so directly relevant, e.g., “I’m losing focus” or “I learned the same thing in high school”. Analyzing the degree to which comments are relevant to the course can help instructors discover the attentive status of learners. Our system classifies comments into two categories, course-related and course-unrelated - using a naïve Bayes classifier.

Content analysis. The content of learners’ comments is diverse and often complex. Some special events are crucial for instructors such as questions, extra opinion and complaints which could directly express the requests or status from learners. Thus, we have adopted instructors’ recommendations by following an existing scheme [MRM13] that divides the content of comments into six types: general conversation, notes, opinion, questions, complaints, and compliments.

Topic analysis and clustering. According to our second design principle, topics summarized from learners’ comments should provide a good overview of such comments, which may reflect the course content, or social activities that interest learners more. Our system applies cosine similarity to group comments into topic clusters and summarize keywords for each cluster.

4.1. Preprocessing

We preprocess comments using a Chinese text-segmentation module, Jieba (https://github.com/fxsjy/jieba), which splits comments into words. After segmentation, we apply stop-word filtering and produce a part-of-speech (POS) tag for each word. Each comment then contains only meaningful words and can be represented as a set of words \( D_i \).

4.2. Emotion, Course-relevance, and Content Analysis

We adopted supervised learning to classify the emotion, relevance, and content of comments. To avoid using a large amount of training data, which would usually be manually labeled, we chose a naïve Bayes classifier to classify comments as it performs well when training data is limited. Each type of analysis is formulated as a two-class classification problem, i.e., positive vs. negative emotion, related vs. unrelated to course content, and a specific type of comment vs. the other types of comments.

Our emotional-analysis approach is modified from that proposed by Hu and Liu [HL04], who used adjectives as opinion-words to mine and summarize customer reviews. They first identified the semantic orientation of each opinion-word by utilizing the adjective synonym set and antonym set in WordNet. They then predicted the
semantic orientation of an opinion-sentence by counting its positive and negative adjectives. If there were more positive adjectives, the semantic orientation of the sentence was classified as positive. We also utilize adjectives for emotion analysis. However, because WordNet does not support Chinese, we obtained the emotional score of adjectives from the training data, and used this score as a feature in combination with naïve Bayes as a classifier to arrive at the overall emotion of a comment.

In the training stage, we counted the occurrences of each adjective in positive and negative comments in the training data. The score of a particular adjective was then defined as \( s = (N_p - N_n) / N \), where \( N_p \) and \( N_n \) are the number of occurrences of the adjective in positive and negative comments, respectively, and \( N \) is the number of occurrences of the adjective in all comments in the training data. After training, the score of each comment to be classified was computed as the average score of all the adjectives it contained: \( \bar{s} = \frac{1}{M} \sum_{k=1}^{M} s_k \), where \( M \) is the number of adjectives in the comment; \( s_k \) is the score of the \( k \)th adjective of the comment obtained from the training process. It should be noted that we only counted those adjectives that appeared in the training comments.

Once the comments’ scores have been computed, we can classify the emotion of each one using naïve Bayes classifier. Specifically, a two-class Gaussian classifier. The decision boundary is obtained by solving \( x \) in the following equation:

\[
\log \frac{P(C_1)p(x|C_1)}{P(C_2)p(x|C_2)} = \log \frac{P(C_1)}{P(C_2)} + \frac{1}{2} \log \frac{\|\Sigma_1\|}{\|\Sigma_2\|} - \frac{1}{2} (x - m_1)^T \Sigma_1^{-1} (x - m_1) - \frac{1}{2} (x - m_2)^T \Sigma_2^{-1} (x - m_2) = 0,
\]

where \( p(x|C_1) \) is a Gaussian distribution function with mean vector \( m_1 \) and covariance matrix \( \Sigma_1 \). As we classify the emotion of comments into positive, neutral, and negative, two classifiers are trained: positive vs. neutral and neutral vs. negative. After classification, we reverse the opinion-orientation of the sentence if there is a negation word close to the adjectives (i.e., word distance less than five), as suggested by Pang and Lee [PL08].

The training and testing procedures of our course-relevance analysis are the same as those of the emotion analysis, except in two particulars. First, in course-relevance analysis, we compute the score of all words, not just adjectives. Second, the relevance score of a word is defined as \( s = N_t / N \), where \( N_t \) is the number of occurrences of the word in the course-related comments, and \( N \) is the occurrence of the word in all training comments. In the testing process, the relevance score of a comment to be classified is defined as the average score of all words in the comment. As with our emotion analysis, words that did not appear in the training process are not counted.

Our content-analysis approach closely parallels course-relevance analysis as described above. For each type of comment, we trained a two-class classifier: target type vs. non-target type. For example, if question is the target type, then all comments that are not questions are treated as the non-target type. We trained six classifiers corresponding to the general conversation, notes, opinion, question, complaint, and compliment types of comments. To resolve the problem that a comment could be classified as multiple types, we count the histogram of each type of comments and assign a multiple-typed comment to a type with fewest comments since the type is more specific.

As more fully explained in the Section 6.1, we collected a total of 1,277 time-anchored comments, representing approximately equal numbers of comments (i.e., 430, 420, and 427) from each of the three online courses we studied. For each course, one-third of the comments were used for training and the others for testing. The average classification rates for the three courses at the test stage ranged from 72% to 90%.

Although the classification accuracy ranged from 72% to 90%, our results of study 2 (Section 6.3) showed that participants did not really notice misclassification, and that their analytics and judgments were not influenced. This clearly shows the promise of applying machine learning techniques to comment analysis and the generation of visualizations. The viability of automatic comment analysis and interactive visual analytics techniques can greatly enhance the effectiveness of MOOCs, and increase the overall impact.

4.3. Topic Analysis

Our course-relevance analysis technique was applied to classify the relevance of each comment. For course-unrelated comments, we adopted k-means clustering to group them. Cosine similarity was used to measure the distance between two comments \( D_i \) and \( D_j \):

\[
\cos(D_i, D_j) = \frac{D_i \cdot D_j}{\|D_i\| \|D_j\|},
\]

where \( D_i \) and \( D_j \) are described by two \( n \)-dimensional vectors computed as follows. For each word in \( D_i \) and \( D_j \), we compute its term frequency-inverse document frequency (tf-idf) value. As the words in \( D_i \) and \( D_j \) may not be exactly the same, we sort the words in \( D_i \) and \( D_j \) and store their tf-idf in the sorted order. We can then form two vectors of equal size by filling zeros to those dimensions that correspond to unmatched words. A high similarity score between two comments indicates that both comments have words in common and that their common words are relatively rare with respect to the rest of the words found in all comments.

We assume that course-related comments are closely related to the course content. Hence, we record the instructor’s scripts with timecode and manually divide each script into multiple segments, each of which is treated as a comment by the instructor and preprocessed using word-segmentation and stop-word filtering in the same manner as for learners’ comments. The preprocessed script comments of the instructor are then used as time-varying templates with which to group course-related comments. We use a similarity function consisting of cosine similarity (Eq. 2) combined with the temporal closeness between the video timecode of the script and that of learners’ comments. Formally, the similarity between a course-related comment \( D_i \) and the instructor’s script-comment \( T_j \) is defined as \( c \cdot \cos(D_i, T_j) + (1 - c) \left[ |t_i - t_j| \right] \), where \( c \) is a weighting constant, and \( t_i \) and \( t_j \) are the timecode of \( D_i \) and \( T_j \), respectively. Each course-related comment is assigned to the instructor’s script-comment that is the most similar to it. In this way, we are able to cluster course-related comments according to the instructor’s script. After topic clustering, the words with tf-idf values larger than a chosen threshold in each topic-cluster are selected as the keywords for each topic.
5. Visualization Method and Implementation

We proposed an interactive visual analytical system that integrates ToPIN and ThemeRiver, with the former providing qualitative analysis of the comments by showing the topics summarized from individual comments as well as the emotional tone and course-relevance of each topic. ThemeRiver is utilized to depict and provide quantitative analysis of the variations in comment types. Please see also the supplemental materials for a demo video of our visualization system and results.

5.1. Visual Encoding of ToPIN

According to design principle 2, we designed ToPIN visualization with two main goals, presenting the evolution of topics and the multi-faceted attributes for each comment and topic. ToPIN depicts the evolution of topics in time-anchored comments as boxes scattered in a two-dimensional space. The x-axis is the time axis and the y-axis displays the values of a user-chosen attribute (Fig. 1). Each box is attached to the time axis by a series of curved lines of various colors, each of which denotes a comment, and all connected to the time axis. The intersection-point of each curved line with the time-axis indicates the timecode in the video at which the comment was left. The density of the curved lines therefore indicates the distribution of comments along the timeline, and their colors represent the comment types, e.g., yellow for general conversation.

The popularity of a topic (measured by the number of comments that mention it) is visually represented by box area. In addition, the shade of a topic box represents the average emotional values of all comments associated with the box, with lighter shades corresponding to more positive emotional values. In other words, ToPIN allows its users to efficiently explore the topics of learners’ discussions about a course, along with the relative importance and attributes of each topic.

5.2. Interface

We implemented our system using Processing (https://processing.org). The interface consists of four views: ThemeRiver, ToPIN, Course Video, and Comment (Fig. 1). The control panel, integrates several user-interface functions. All widgets can be resized and moved according to users’ preferences. To prevent the visual cluttering problem and enhance the efficiency of using interface, our system supports a set of interactive functions as below.

Highlighting. When a user clicks a topic box in the ToPIN view, the topic box and all comment curves connected to that box are highlighted and the keywords describing the topic are listed in a pop-up box. Similarly, in the ThemeRiver view, clicking on a flow will highlight that flow, which represents a specific type of comments. In addition, the Comment view lists only those comments associated with the highlighted topic/flow, so that users can focus on the relation between topic content and the timeline distribution of the associated comments. Users can thus intuitively browse a specific type of comment that they are interested in.

Display filters. A user can choose to see all, or only certain types of comments by clicking the six boxes in Content Type Selection. This could help users to find out which part of a course video prompted the most questions, or if learners took more notes on the most important part of a lecture, and so forth. The ToPIN, ThemeRiver, and Comment views are updated simultaneously when the user makes changes to Content Type Selection. We also designed a Comment Count Filter, operated via a slider bar, which allows users to screen out less popular topics, i.e., topics with comment counts lower than a threshold amount set by the user.

Synchronous time bars. To help users establish temporal associations between different visualization views and video content, the time bars in the course video, ToPIN, ThemeRiver, and Comment views are synchronously updated. In other words, when the user scrolls a time bar in any of these views, the time bars in the other views will be updated to the same time instant. This ensures that users are able to accurately explore the contextual relations among comments, course video, and visualizations.

Plot selection. In the control panel, the user can choose to see only ThemeRiver, only ToPIN, or both of them by clicking a visualization mode. In addition, the user can plot emotional valence or course relevance on the y-axis. If course relevance is chosen, a topic box located in the upper part of the screen will indicate the relevance level of its contents to the course content is high.

6. Evaluation

We conducted two user studies. The first one aims to ascertain whether instructors could use our system to discover and analyze feedback from learners effectively. The goal of the second study was to know whether educational evaluators who are not familiar with videos could use the system to effectively get the trends of learners’ status and find out the strengths and drawbacks in
lecture videos. In particular, the second study compared how the educational evaluators explored learners’ comments with and without visualization to evaluate the efficiency of our interface. Additionally, the second study incorporated a feasibility study of our comment-analysis approach.

6.1. Data Collection
The time-anchored comments in this experiment were collected in a previous study [LLC*15]. A total of 50 online-learner participants (21 females, 29 males, M = 20.6 years old) were recruited to watch the course videos and leave comments. Three online courses were selected from the OCW and Coursera of major universities in Taiwan. One video clip with a length of approximately 15 minutes was selected from each course, including “Reward and Addiction” from a neural-science course “Introduction to Economics” from an economics course, and “Understanding the Greek Philosophy” from a philosophy course. The number of comments collected from the neural science, economics, and philosophy courses were 430, 420, and 427, respectively.

We convened a focus group consisting of two researchers who coded the emotional tone, course relevance, and content type of each comment, and grouped comments into topic clusters. When the two human-coding results reached 90% similarity, one of them was adopted as the ground truth for comment classification and analysis.

Fig. 2 shows our visualization of the learners’ comments in the philosophy video. From this, it can be readily observed that there were two major course-unrelated topics characterized by negative emotion, and that the comments associated with these two topics were spread along almost the entire timeline. Many of the curved lines associated with the participants’ comment were red (complaint) or yellow (general conversation), indicating that our pool of learners had many complaints and could frequently be distracted. The keywords of these two topics were “abstract and hard to understand” and “fast speed of changing the course slides”, which reveal that the learners were complaining that the course content was too difficult and that the teaching style was unhelpful. Fig. 3 shows our visualization of the economics video. It reveals that learners kept taking notes (pink flow) throughout the lecture and there are several topic boxes containing compliments (green curves). These suggest that most learners were concentrated and satisfied with the lecture. Please see the supplemental materials for the visualization of the neural-science video.

6.2. Study 1: Supporting Instructors’ Analytics
We invited the instructors of the Neural Science (P1) and Economics (P2) courses to evaluate our system.

6.2.1. Task design
We focused on testing the overall usability of our system and how well it assisted the instructors in exploring learners’ feedback. In the first task, the instructors were asked to observe the trends in learners’ participation in their courses. In the second task, the instructors were asked to tell us which course content induced heated discussion among the learners. In the third task, we allowed the instructors use our system to freely explore learners’ comments. After completing these tasks, the instructors were interviewed and asked to explain their strategies for using our system.

6.2.2. Procedure

Figure 2: (a) ThemeRiver and (b) ToPIN visualizations of learners’ time-anchored comments on the online video of a philosophy course. It can be observed that comments related to the video content tend to be emotionally more positive than the unrelated ones.
We first introduced our visualization system to each instructor and conducted pre-interviews with them. The purpose of the pre-interview process was to understand the instructors’ experience of teaching on online courses and what information they wanted to obtain from learners’ feedback. After each pre-interview, we demonstrated our interface using the comments on the philosophy course, which were not subsequently used in the study itself. Then, we gave each of them a practice task involving the philosophy course. If they completed this practice task successfully, it would indicate that they would be able to perform the three tasks we had designed as the main study. None of the instructors failed to complete the practice task. In each main task, we conduct the think aloud method, then record the behaviors of and thoughts reported from instructors. Once the main study was completed, we interviewed each instructor again to learn about their user experiences and what they had found out about the students’ course experiences.

6.2.3. Results and discussion

Task 1: Observing the trends of learners’ participation in courses. On their own initiative, the instructors developed the following strategy. They tended to first observe the flow of the entire ThemeRiver to see where the most or fewest comments had been left. They then selected different content types to observe the flow of each. P1 found that most comments appeared at the beginning of the video; however, most of these comments consisted of general conversation. P2 reported that the learners maintained heated discussion throughout the entire video, and many topics were characterized by positive emotions. Both P1 and P2 concluded that the learners were engaged by the course videos and had heated discussion on diverse topics including both course-related comments (e.g., notes) and course-unrelated ones (e.g., general conversation).

Task 2: What course content stimulated learners’ most active discussions? Both P1 and P2 chose to set a higher threshold in the comment count filter to screen out less popular topics. To locate the most active discussions, they both selected opinion content type in the control panel and used ThemeRiver to observe the flow of opinion and question comments. Next, according to the peaks of flows in ThemeRiver, they both switched to ToPIN to check the topics at the time points of these peaks. P1 found that there were many large topic boxes about dopamine, which was one of the most important subjects of this course. P2 found that the learners had numerous questions about the Demand Law. P2 expressed his surprise about learners’ feedback on the Demand Law and said he was considering modifying his teaching materials on this subject.

Task 3: Freely exploring comments. P1 mainly used the ToPIN view and the comment-count filter to identify major topics. He generally focused on the question-type comments. Unsurprisingly, given that P1 had reported his primary concern as being whether the learners would be able to understand the course content, he browsed all the topics containing question-type comments and then looked through all the questions on each topic. He found some questions from learners that revealed their misunderstandings of terminologies and definitions, as well as confusion that had been caused by his unclear pronunciation. After browsing learners’ questions, P1 reported that he would improve the clarity of his lecture.

P2 first observed the flows of all comment types in ThemeRiver, and then viewed the peaks of each flow. Browsing the flow of compliments, P2 found a high peak at the end of the video. He then switched to ToPIN, and found there were some topic boxes around the time-point of the peak containing many compliments about his interesting way of teaching and the excellent examples he had used.
to illustrate the difference between superior and inferior goods. P2 was excited to find how much the learners liked the way he had designed this part of the course, and said he cared very much about whether the learners would accept his teaching methods.

**Interview results.** In their pre-interviews, we asked P1 and P2 instructors about their experience of using visualizations when teaching online courses, and both replied that they had only used pie- and bar charts to visualize learners’ behavioral data, such as the number of times they had logged in or when they had downloaded the course materials. P1 and P2 both stated that visualizing feedback and comments from learners was necessary and could help them improve their teaching continuously. They anticipated that our system could provide them with very useful suggestion through the learners’ comments. As P1 put it, “I really want to know the positive and negative suggestions from learners. I also want to know the shortcomings of my online lecture. I hope I can find some concrete feedback to improve my course content.”

Based on the post-study interview from instructors, they both reported that the flows of ThemeRiver help them easily learn the trends of different content-type comments. For ToPIN, the summary and keywords of each topic allow the instructors to quickly comprehend the contents of heated discussion and the reason why learners left negative comments. P1 also reported he’s interested in every topic clustered from learners’ comments, and P2 said he would be curious about each larger topic box. Both instructors prefer the content type filtering most because it assist them focusing on question, opinion and notes comments which usually reflect the learners’ level of interest in the course. Last, the comments characterized by negative emotion strongly attracted the instructors’ attention as they potentially reveal the specific reasons that some learners complained about the course.

**Discussion.** In Study 1, temporal visualization allowed the instructors to interpret the features presented in the course video quickly and accurately. Both ToPIN and ThemeRiver were easy to learn for P1 and P2. First, this study showed that the instructors used ThemeRiver to assess the overall situation of learning with the course video (i.e., the global pattern). The instructors indicated that the flows of different colors in ThemeRiver were very suitable for illustrating the trends of learners’ comments. Second, they used ToPIN to observe the details of learners’ comments (i.e., the local details). Because we presented the attributes of comments and clustered them into different topics, it gives users an easy way to understand the features of each topic, e.g., the crucial event and the learners’ satisfaction of certain course contents. After learning the trends of comments by ThemeRiver, instructors switched to ToPIN to get the detailed information. Third, Our analysis methods, particularly content analysis, effectively help instructors retrieve the information they are highly interested and gain novel insight into the information. For example, instructors considered the notes could represent the attentive status of students, the question and opinion the heated discussing patterns.

Both instructors said that one of the greatest strengths of our visualization was that it showed temporal patterns of topics, comments, and course video integratively. Instructors could therefore observe the segments of greatest interest to them on the video timeline, and gain the insight from the context of the comments directly.

“Because the topic keywords of the learners’ comments are displayed, I can easily retrieve the corresponding course topics.” (P1)

“By using the system, I can identify the reasons that the negative emotions or some specific responses of learners were generated. After I conclude the reason, I can come up with some solutions to prevent the problem from happening again.” (P2)

This verified that our system can help instructors to understand which content interests learners the most and least (among other matters) and locate special events or responses from learners’ comments through visual analysis.

### 6.3. Study 2: Supporting Educational Evaluators’ Analytics

The goal of this second study was to establish whether our system could facilitate valid and useful evaluation when the users are not familiar with lecture video, and particularly when the comments analysis is done automatically with machine learning rather than manually by human coders. In study 2, we conducted a within-subject experiment and invited educational evaluators to assess the learning quality and states of online learners by ten-evaluating tasks both with and without using our visualization system. We recruited eight participants (four females), with an average age of 23.7 years. All participants had a background in education, instructional design, or learning science. None of them had prior knowledge of our video clips.

**6.3.1. Task design**

We designed ten tasks based on previous research in educational evaluation [Kir97, Cro13] and feedback from the instructors in Study 1. These tasks, which are listed in full in the supplemental materials, can be categorized into three groups.

In the first group, we chose some characteristics that instructors concerned about and asked the participants to identify which course contents meet these characteristics from learners’ feedback and comments, e.g., what video content the learners found most difficult, interesting, or unsatisfying (Task T1, T3, T4, T5). In the second group, we asked the participants to locate particular time-points at which the given specific events occurred, e.g., when learners paid more/less attention or engaged in more discussion (T2, T6, T7). The third group of tasks required the participants to provide general comments on the teaching skills of the instructors (as ascertained from the video and learners’ time-anchored comments) and their suggestions for improvements to the course video (T8, T9, T10).

**6.3.2. Interview design**

We also conducted a two-part interview with each participant. The first part (Q1, Q2) took place after a participant completed each of the two use conditions: without using the visualization (i.e., baseline that resembles the current practice) and with the visualizations. Q1 and Q2 basically asked participants about whether they considered any task very difficult and whether they make any additional discoveries not related to the required tasks. The second part of the interview (Q3- Q6) was conducted after a participant had completed both conditions. Q3-Q6 emphasized on the comparison of their experiences with using versus not using visualizations in their analytics and their confidence level about their analytics in
the two conditions. The interview questions are listed in the supplemental materials.

6.3.3. Procedure

We first introduced our system and explained the overall procedure. Then, we gave each participant a five-minute practice trial of the interface using the philosophy course video (which again would not be used in the main study). Immediately following the practice trial, the participant commenced the main-study tasks. These were divided into two sessions: one without, and one with, our visualizations. In the baseline (no-visualization) condition, we displayed our automatic comment analysis results, including the course relevance, emotional tone, and content type of each comment, using symbols or colored squares in the command view, but did not display the visualization interface. Users could search the comments and use the content type selection.

To avoid ordering effects, we counterbalanced the order of the sessions with and without visualizations, i.e., with half of the participants performing the session with visualization first. Each participant also evaluated two different course videos (neural science and economics), one in each of their two sessions. During the study, we asked each participant to use our interface to complete an evaluation questionnaire regarding the educational video. After the first session, we interviewed the participant about the strategy they applied to find out the answers. In the end, we interviewed the participants again to record their use experiences and asked them to compare the differences between the two sessions.

6.3.4. Results and discussion

User behaviors. In the baseline condition (without visualization), we observed that all participants had to browse/skim almost all the comments to complete the tasks and then summarize the findings by themselves. From the interview, we found that they tend to first select a specific content type, then view the comments individually to obtain detailed information. Some also counted the number of comments containing specific keywords relevant to the most active discussions of certain topics. The participants reported that it was difficult for them to complete tasks for uncovering specific attributes and states of the course content (e.g., level of difficulty, students’ degrees of confidence and satisfaction). “I can find the specific topic by viewing the bigger topic box. But in the previous condition (baseline), I have to browse all the comments and then identify a specific topic. It was more time-consuming…” (S7)

In the condition with visualizations, five of the eight participants adopted a strategy similar to that of the instructor-participants in Study 1. At first, they used the display filter to select content type and applied ThemeRiver to observe the flow of the selected content type. After learning the trends of comments, they switched to ToPIN to further explore topic content, then shifted to the comment view for the details of comments. Of the other three participants, two (S1, S6) mainly used ToPIN and content-type selection to complete the tasks, while the other (S2) did not use visualizations at all. S2 reported that she had not used any information visualization tools before, and was more comfortable with browsing comments directly even if visualizations were provided.

With visualization, participants more efficiently finished tasks than without visualization. We analyze the time participants spent on ten analytical tasks we requested. In terms of descriptive statistics, the average and standard deviation of time required for task completion in the with-visualization condition are 36.6 and 14.8 (mins); which are 46.2 and 23.1 (mins) in the baseline condition of no visualization. S8 withdrew the tasks in the baseline condition because he reported that it’s too hard to finish tasks by using the regular baseline interface. The results shows that our visualization enhances the efficiency to observing time-anchored comments and obtaining the insights in contrast to baseline methods.

Participants had better consistency of answers and greater breadth of findings when using visualization. For T1-T7, we analyze the consistency and breadth of the insights generated by our participants. In the condition with visualizations, participants’ answers appeared to be more consistent with one another than they were in the no-visualization condition. With regard to the neural-science course, S1, S5, and S6 provided similar analytical results. For example, they all provided keywords - “dopamine decrease”, “midbrain”, “thrill”, and “emotion” - to T5, which required them to. However, in the baseline condition, participants reported a variety of different answers to the same task, e.g., “reward system” (S3), “the definition of dopamine” (S4), “the function of dopamine, video, Prozac” (S7), and “terminologies” (S8).

Similarly, for task T3 in the economics course, S4, S7 and S8, all consistently identified that the “opening of the course” is the part where learners showed the greatest interest. However, participants in the baseline condition, including S1, S2, S5 and S6, failed to show consistency, identifying all different parts of the video as where the learners were the most interested in, such as “inferior goods”(S1), “slides of the instructor” (S2), “opening of the course”(S7), “the three kinds of goods”(S8).

For the breadth of insights, we counted the numbers of findings from each participant in T1 and T3-T5. The averages number of findings generated were 10.5 and 7 for neural-science course video when visualization is available and unavailable respectively; 9.3 and 8.5 for economics course video in with-visualization condition and baseline condition, respectively. Participants generally discover 2.1 more findings when using our visualization support.

Participants felt more confident when using visualizations. In their interviews, six participants stated that they were more confident in the answers they gave in the with-visualization condition. They felt that they could trust the statistics/analysis of comment counts and topic keywords that ToPIN and ThemeRiver provided, and that ToPIN and ThemeRiver helped them reduce the burden of analyzing and arranging all the comments themselves. Likewise, these six participants mentioned the uncertainty and lack of confidence caused by human factors (e.g., fatigue) if no visualizations were provided.

However, two participants (S2, S6) preferred to view the comments directly. They did not trust ToPIN and ThemeRiver, because they had found the system had provided some wrong categorizations of content type, emotion, and relevance. Nevertheless, we found that S6’s answers to all tasks were more accurate in the with-visualization condition than in the baseline condition. This might suggest that ToPIN and ThemeRiver helped the participant complete the tasks even though she did not actively use them.
7. Design Implications

Time-anchored commenting visualizations can provide multifaceted analysis of learners’ feedback effectively and efficiently. Because conversational data such as comments reflect the diversity and complexity of humans’ textual expression, computational supports, such as visualizations, are required for deep exploration of learners’ comments on a large scale. Our system helped the instructors quickly focus on course-related discussions, which is desirable for MOOC instructors [SMHF14].

Quantitative and qualitative visualizations of comments could enhance understanding of online learners’ tastes of learning. Our system integrates ThemeRiver and ToPIN, which support quantitative and qualitative visualization, respectively. Both the instructors and three-fourths of the education researchers in our sample took advantage of both types of visualization, switching between them to explore the learners’ comments efficiently and effectively. This indicated that quantitative and qualitative information could offer the participants a great range of information useful for inferring learners’ tastes of learning. Many participants also mentioned that the integrated display of the information about learners and course content provided a comprehensive view of the context of commenting. This finding also supported our design principle 2.

Time-anchored commenting offers a useful channel for monitoring online learners’ learning experiences. Time-anchored commenting was originally proposed as a means of improving learners’ social interactivity and engagement [LLC15]. In the current study, feedback from course instructors and educational evaluators showed that time-anchored commenting could also be a useful mechanism and source of information for understanding students’ learning status and experiences, and especially their interaction with course content.

8. Limitation and future work

In this work, the visualization design and encoding and text processing methods we applied are the prototyping version to visualize both the quality and quantity of time-anchored comments. With the positive results for our initial designs, we believe, in the future, the replacement by more sophisticated techniques will cause more impact on MOOCs. In addition, visual cluttering is a challenging issue for handling large scale learners. In our current tool, we designed several interactive filtering functions to reduce the clutter problem. In the future, we will try to apply other techniques, e.g., hierarchically presenting the topic boxes by the numbers of comments. Beside, the comment data we used were collected in a lab setting [LLC15], in which the number of learners is smaller than that present in a regular MOOC course. Finally, we would like to utilize conversation visualizations like [FZCQ17] to help instructors trace conversations among individual learners and discover those learners who need help.

9. Conclusion

This paper presents a comment analysis and visualization system for online education videos. The two evaluation studies participated by the course instructors and educational evaluators validated the efficacy of our system. These two evaluation studies further suggested that time-anchored commenting can be a very useful channel for monitoring online learners’ feedback and with the assistance of our visualization techniques, the course instructors and educational evaluators can efficiently gain insights into students’ learning status and quality of learning.

References


[Cro13] CROSS S.: Evaluation of the old mooc curriculum design course: participant perspectives, expectations and experiences. 8


[QCL∗15] Qing C., Chen Y., Liu D., Shi C., Wu Y., Qu H.: Peakvisor: Visual analytics of peaks in video clickstreams from massive open online courses. 3


[SFCQ14] Shi C., Fu S., Chen Q., Qu H.: VisMOOC: Visualizing video clickstream data from massive open online courses. In VAST (2014), pp. 277–278. 1, 2, 3


