DropoutSeer: Visualizing Learning Patterns in Massive Open Online Courses for Dropout Reasoning and Prediction

Yuanzhe Chen*[1], Qing Chen†[1], Mingqian Zhao‡[1], Sebastien Boyer§[2], Kalyan Veeramachaneni¶[2], and Huamin Qu, *Member, IEEE*‖[1]

[1]Hong Kong University of Science and Technology
[2]Massachusetts Institute of Technology

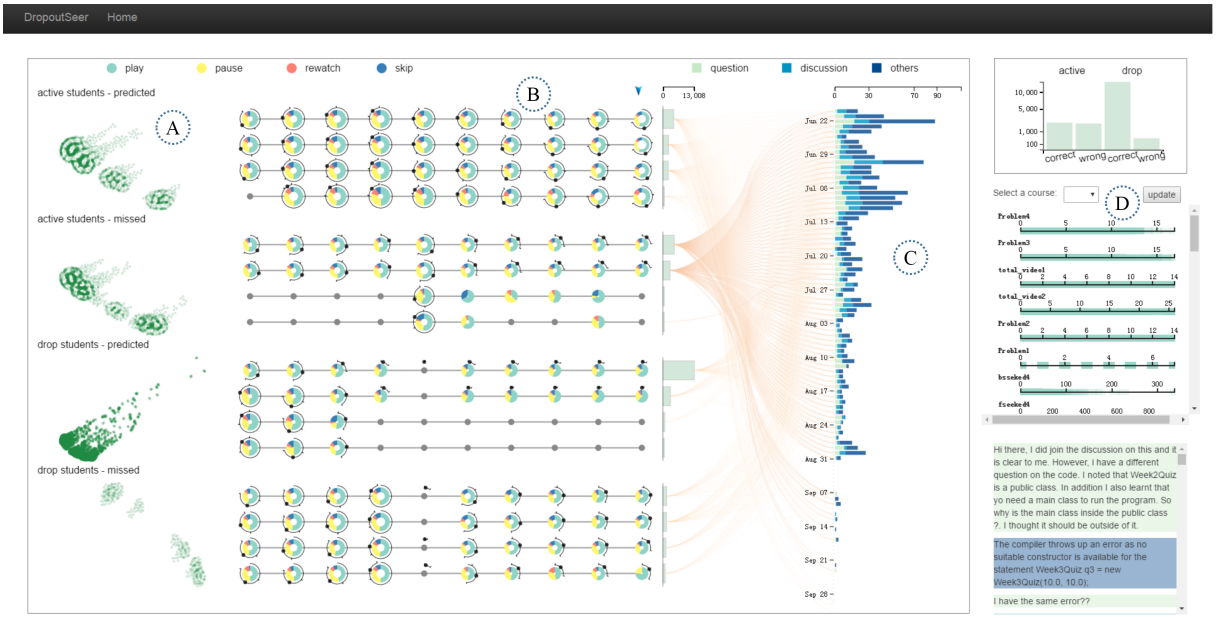

Figure 1: The figure shows the DropoutSeer system for the analysis of a *JAVA* programming course: (A) shows the learner groups clustered by their learning activity. (B) displays the clickstream behavior and the assignment performance of different learner groups along the timeline. (C) presents the posts of learners on the course forum. (D) lists general information including the overall distribution at the top, the dashboard in the middle, and the forum content at the bottom.

Abstract

Aiming at massive participation and open access education, Massive Open Online Courses (MOOCs) have attracted millions of learners over the past few years. However, the high dropout rate of learners is considered to be one of the most crucial factors that may hinder the development of MOOCs. To tackle this problem, statistical models have been developed to predict dropout behavior based on learner activity logs. Although predictive models can foresee the dropout behavior, it is still difficult for users to understand the reasons behind the predicted results and further design interventions to prevent dropout. In addition, with a better understanding of dropout, researchers in the area of predictive modeling in turn can improve the models. In this paper, we introduce DropoutSeer, a visual analytics system which not only helps instructors and education experts understand the reasons for dropout, but also allows researchers to identify crucial features which can further improve the performance of the models. Both the heterogeneous data extracted from three different kinds of learner activity logs (i.e., clickstream, forum posts and assignment records) and the predicted results are visualized in the proposed system. Case studies and expert interviews have been conducted to demonstrate the usefulness and effectiveness of DropoutSeer.

Index Terms: Visualization in education, machine learning, time series data, design studies

*ychench@ust.hk
†qchenah@ust.hk
‡mzhaoad@ust.hk
§sebboyer@csail.mit.edu
¶kalyan@csail.mit.edu
‖huamin@cse.ust.hk

1 Introduction

MOOCs have become increasingly popular since 2012, during which the three major MOOC platforms have been launched. Quite a large number of universities and educational institutions have spent considerable time and effort in developing and delivering high-

quality courses. However, many problems regarding MOOCs and their effectiveness remain unsolved. One of the most crucial issues which have been widely debated is the high dropout rate. Although more than 10000 registrants often enroll for a single MOOC course, the dropout rate is usually higher than 70%, and can even reach 90% [4, 25, 32]. This problem has aroused much attention from both the researchers and the public.

To tackle this problem, researchers in the field of machine learning have built various models to predict dropout. The benefit of these predictive models is mainly two-fold. First, predictive models not only calculate the dropout probability of different learners but also identify crucial factors correlating to it. Using these factors, instructors are then able to design interventions to keep or even enhance the engagement of learners. For instance, with the help from instructors or teaching assistants, learners who intend to finish the course but still encounter some unsolvable problems might continue with their studies. Second, dropout prediction facilitates the understanding and classification of various motivations and behavior of online learning. Usually an online course attracts learners from diverse educational and cultural backgrounds with different motivations towards the course. By analyzing and predicting learners' actions, predictive models can provide researchers with clearer relationships between learners intentions and their learning activity.

Although predictive models are powerful in terms of identifying learners who may drop the course, there are still huge gaps between the predicted results and the dropout reasons. On the one hand, dropout analysis involves MOOC data, which in nature are large scale, heterogeneous and temporally evolved. Also, the data contain noises and only a small number of learners remain active throughout the course period. Therefore, it is extremely difficult for instructors and education analysts to associate the likelihood of dropout with the learning activity. For example, if a learner rewatches a certain video segment for many times right before finishing a related assignment problem, his/her assignment performance may be enhanced if the relevant reviewed content is clearly conveyed, thereby motivating the learner to continue his/her course. By contrast, if a learner asks questions on the forum but receive no response or no valuable feedback, he/she may give up the assignment or even quit the course very soon. On the other hand, the reasons of dropout are diverse and highly personalized. Even for experienced instructors, they still require tools to help them associate the learning activity with the potential reasons of dropout. Visual analytics methods and techniques are suitable for tasks that analysis of large amounts of information is required and can not be solely solved by computational methods or human effort [35, 21]. Therefore, we believe that a visualization system can be beneficial for solving the dropout problem, and this observation initially motivates our work.

By correlating the predicted results from models with the learning activity, not only can instructors better design a MOOC course, but also can machine learning researchers construct more effective predictive models. For instance, two recent works [20, 34] claim that feature ideation is a critical step in the model building process to build a highly accurate model. However, given the complexity of MOOC data, less effective features become the bottleneck in many existing models. By exploring the data using a visualization system, machine learning researchers can obtain feedback on their models and better understand the data even with little domain knowledge. In this way, more accurate models can be built.

To address the aforementioned problems, we present a visualization system, DropoutSeer, which allows both instructors and researchers in predictive modeling to better understand the reasons for the dropout behavior through joint analysis of video watching behavior, assignment performance and forum discussion. We follow a user-centered design process and involve both course instructors and machine learning researchers at each stage of the iterative system development. During this process, we realize that normally instructors do not have much knowledge in machine learning nor a strong mathematical background. Accordingly, intuitive visual designs must be proposed to users with relevant information. Moreover, upon the prediction results, users also need to classify them into meaningful subgroups so as to find regular and irregular patterns, and identify the typical learner clusters and outliers. Therefore, the system should be both intuitive and interactive for end users.

DropoutSeer contains four linked views: 1) a cluster view presents the clustered learner groups; 2) a timeline view illustrates the clickstream behavior and the assignment performance of different learner groups; 3) a flow view links the posts of learners with the timeline view; and 4) a dashboard view lists some general information and allows various filtering. We also carefully design a novel diagram for temporal pattern detection and smooth transitions over different views. To evaluate our system, we conduct both case studies and expert interviews with five end users. The results prove the usefulness and effectiveness of our system.

The major contributions of this paper are as follows:

- A visualization system which integrates four linked visual designs to enable analysts to identify learning patterns related to dropout behavior at multiple scales.
- A novel temporal data visualization design which uncovers the learners' detailed learning activity.
- Case studies with real datasets and expert interviews with domain experts to demonstrate how our system can help instructors and machine learning researchers to analyze the reasons for dropout behavior.

2 Related Work

This section presents current research on dropout prediction, visualization in MOOCs, time-series visualization and predictive model visualization.

2.1 Dropout Prediction

Many works have tried to describe the dropout prediction to a binary classification problem. We first review the features used for this task, and then introduce the machine learning models adopted.

The common features used in existing works are learners activity logs of accessing different parts of a course such as the videos, the course forum and course wiki [2, 20, 29, 34]. The activity logs can be first separated into different time periods (e.g., by week) and then concatenated and represented by their statistical features [2, 20, 29, 34]. Sinha et al. [32, 33] further considered the students activities as sequential data. They also formed a directed graph to represent the sequential structure [33], and extracted features based on several typical short activity sequences rather than single activities [32].

With the extracted features, many classical machine learning models have been tested, including Support Vector Machine (SVM) [2, 20, 33], Decision Tree [29], and Logistic Regression [4, 34]. The accuracy of most existing works are over 80 percent. However, these models are designed to predict dropout based on training data extracted from the same course, which means they are not suitable for predicting an ongoing course in real-time. On the other hand, Boyer et al. [4] applied transfer learning models to predict dropout of an ongoing course with training data from another course. This model is adaptive to the difference between the two courses. Nevertheless, we believe such difference is still difficult to eliminate in real use. Therefore, we hope the system presented in this paper can help researchers in data mining to obtain more adaptive models.

2.2 MOOC Visualization

Compared with traditional education records, the data from MOOCs have much finer granularity and contain new pieces of information. These characteristics require lots of visual encodings to assist users in finding patterns. Some studies have used basic visualizations such

as bar charts and line charts to reveal general patterns or the learner distribution. For example, scatterplots were used to suggest that longer videos and rewatching learners often exhibit higher dropout rates [19]. In another example [11], some standard visualizations such as heatmaps and stacked bar charts have also been implemented to present aggregate statistics such as the ratio of the number of certificate achievers to the number of registrants. Others have involved more advanced visual analytics methods to show the relationship among different learner communities or transitions among various actions. Brown et al. presented forum interaction networks and clustered student communities with node-link diagrams [5]. Huang et al. visualized student assignment submission networks with node size representing the number of submissions and color corresponding to the test performance [17]. Coffrin et al. [9] designed a state transition diagram to describe learners' access transitions among different videos and assignments.

Recently, more comprehensive visual analytics systems have been developed. For example, VisMOOC [30] designed a seek diagram for analyzing video clickstream, along with standard stacked graphs. The system aligned the visualizations along the video timeline with the corresponding video, so that users could do content-based analysis. Later on, another work called PeakVizor [8] was proposed to visually analyze the interaction peaks in MOOC video clickstream with some complex visualizations designs such as glyph, flow map, and parallel coordinates. These works provide multiple interactive functions such as sorting, zooming, clustering, and users could filter among different views to find patterns from various perspectives. However, few of these studies have focused on dropout in particular, thus an integrated visual analytics system targeted on dropout is required for users to understand the reasons behind dropout behavior.

2.3 Time-series Visualization

There have been a large variety of visualization techniques for analyzing time-series data, which have been summarized in several comprehensive surveys [1, 3, 24]. The most prevalent approach to represent time is using a horizontal axis [24, 26]. Each attribute of the time-series can be further illustrated using a stacked graph layout [15]. Besides, different visual metaphors were adopted to encode certain types of temporal data. For example, Van et al. [36, 37] used a calendar display to visualize time-series which aggregated on daily, weekly or monthly basis. Dragicevic and Huot [13] used a spiral layout to represent periodic temporal data.

Besides the general techniques for time-series visualization, there have also been various visualization techniques proposed to be domain-specific tools. For instance, Wu et al. [38] proposed a glyph-based technique to visualize dynamic egocentric network. Chen et al. [7] presented the Criminal Activities Network (CAN) to extract, visualize and analyze criminal relationships in the field of law enforcement. In text data visualization, several works [10, 12, 22, 31] have been proposed based on the Streamgraph design for large-scale corpora analysis. To analyze social media data, researchers have also proposed visualization tools for detecting abnormal events [6] and analyzing how multiple topics compete with each other on social media [39]. In this paper, we extend the technique proposed in [38] to visualize the detailed learning processes of MOOC data.

3 Problem Characterization

As shown in Fig. 2, the DropoutSeer system comprises three components, namely, 1) the data manager, which cleans and preprocesses the raw data and stores the three learner activity records into our database; 2) the data model, which contains the predictive model and clustering/classification methods to greatly facilitate the process of pattern discovery; 3) the visual analysis component. This section describes the data manager component and the predictive model, whereas the following section describes the visual designs in detail.

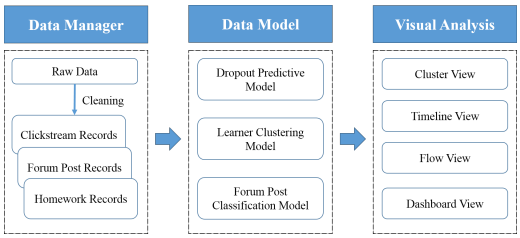

Figure 2: The overview of DropoutSeer system. (a) The data manager cleans and processes the raw data and stores the data in database. (b) the data model incorporates three analytical methods to further process the activity records. (c) the visualization system includes four linked views for users to perform visual analysis interactively.

3.1 Data Abstraction

Provided by MOOC platforms such as Coursera and edX, learners activity log mainly contains three types of data: records of clickstream in course videos, discussion posts on course forums, and grading information for each assignment set. The records of clickstream can be further categorized into play, pause, seek (i.e., jump from one timestamp to another), stalled, ratechange and error. Each record is associated with a timestamp to indicate when the activity has been triggered. These three types of data are hereafter referred to as clickstream, post and assignment. Table 1 shows some general statistics of the log data for the two courses (denoted by *NCH* and *JAVA*) that will be analyzed in our case studies.

Before being input into the predictive model and the visualization system, the raw data need to be cleaned and preprocessed. For example, a play event is generated when a learner initiates a pause or seek action. Therefore, we need to find all such pairs of records and label each pair either as a pause or a seek action. In the raw data, we also find that an individual learner can generate hundreds of clicks within a second, which may be attributed to system errors. Therefore, we remove this kind of records. Furthermore, some statistics need to be inferred from the raw data. For instance, if we want to split the seek action into skip (i.e., seek forward) and rewatch (i.e., seek backward) actions, then we need to infer this action by looking at the click sequence of a learner.

After preprocessing the data, we perform a preliminary analysis of the data with our end users. Three properties of the learners log data are summarized below which guide our visual design.

Bias of inactive learners: Based on the data, we find nearly half of the learners initiates no records at all. For the rest of the learners, the majority of them only remain active for one week. These learners cannot be directly filtered since users prefer a general picture of the dropout rate, but they can hardly generate meaningful patterns for the reasons of dropout. Therefore, we need to keep inactive learners and highlights the active learners in our visual design.

Sparsity of posts: As shown in Table 1, the clickstream and assignment records significantly outnumber the post records. The interviews with our end users show that content of forum posts can accurately reflect the attitudes of learners. Therefore, clickstream and assignment are more suitable for end users to generate hypotheses, which can be validated by the corresponding posts. We also need to consider this difference in our visual design.

Uncertainty of dropout definition: By interviewing our end users, we also find that the definition of dropout may vary among different users. Instructors are usually more interested in whether a learner will finish a course (e.g., receive a certificate or pass the final exam). Other than this, our end users also get interested in when a learner will drop the course or whether a learner will be active in a particular week. Therefore, our visual design must be general in terms of dropout definition.

Table 1: General statistics of MOOC courses

Course	*JAVA*	*NCH*
Platform	Edx	Coursera
# of learners	18832	15232
# of weeks	10	9
# of videos	122	17
# of clickstream records	3197422	1204947
# of assignment submission	239535	15482
# of post	13289	4683

3.2 Task Analysis

Through the data abstraction, Several important properties of the data have been identified. Based on the understanding of the data, we have conducted three rounds of interviews with our end users to characterize the requirements, which are summarized as the following. While all the five tasks listed below are important for helping instructors understand the dropout reasons, only T2, T3 and T4 are relevant to machine learning researchers.

T1. What is the general dropout distribution? In general, our users are interested in identifying some typical dropout learner groups, and the general distribution of dropout and non-dropout learners in both predicted results and actual results.

T2. Are there any dropout learner groups and what are the factors that affect the different groups? Grouping is one of the most interested topics from our previous interviews. Since there are a variety of reasons for dropouts, users also wonder whether there exist some typical dropout learner groups. Also, we have extracted a number of features for dropout prediction. Based on the predicted results, users want to know which feature affects the predictive results more than the others, and which feature is useful to analyze the reasons behind dropout.

T3. What are the learning patterns demonstrated by different learner groups? After grouping learners based on predictive features, users demand for more detailed information about how learners behave throughout the whole course, for instance, whether their temporal behaviors such as video watching and interaction with videos relate to their assignment performances. Some active learners maintain a steady schedule and work hard every week yet cannot obtain a high score, whereas others stay active in merely several weeks but at last turn out to obtain high grades.

T4. How do forum posts correlate to a selected learner group? The discussion in MOOC forums is a principal approach for learners to interact with peers, teaching assistants and instructors. Such forums are often regarded as asynchronous discussion groups. The forum posts of learners within a selected learner group may greatly vary and cover a diverse range of discussion topics. Meanwhile, post records also include temporal information that may reveal some hidden patterns correlated with video watching behavior. Therefore, we need to design a view for the post records to be capable of further explaining or validating the patterns found in clickstream and assignment data of a selected learner group.

T5. What are the learning patterns of a particular learner? After studying grouping behavior, in a pilot interview, users sometimes find abnormal posts on the forum from a small number of learners who may drop the course in a very short time. Therefore, users want to conduct more detailed analysis on an individual level, and reason some special dropout cases with video watching, forum posts and assignment behavior. If joint analysis with coordinated timeline is provided, this would help them identify some typical dropout intention and behavior, which would benefit future adjustments on course materials or special aids for those learners.

3.3 Predictive Model

In addition to the insights derived in the sections above, providing high-level metrics relevant to users' need is our goal here. To do so, we adopt a machine learning point of view on the data and construct probability estimation of the dropout likelihood for each learner.

3.3.1 Dropout Prediction Problems

In this section, we give a formal definition of dropout for the model. We say that a learner drops out at week W if at week W the learner has not left any activity records.Given a particular course we can define many different prediction problems. In the first week, one may want to predict learners who will remain at the end of the course. For different purposes, one may also want to predict learners who are likely to remain in the class in the week right after the prediction being made. Therefore, we call dropout prediction problem a tuple (w_b, w_d) using behavioral data from week w_b to predict which learners will remain in the course in the prediction week w_d.

3.3.2 Learning a Predictive Model

We are now left with: on the one hand a completed course through which we know both the behavior of learners and the resulting outcomes (dropout status during all the different weeks), and on the other hand the dropout likelihood we need to estimate. To produce these estimates, we test the three most common classification algorithms used by the machine learning community. We give the name and a short description of the high-level idea behind each of them.

- We train a Logistic Regression [23] classifier which tries to find the optimal weights w to use on each behavioral feature so that the score (weighted average of the feature values) will enable us to discriminate between learners very likely to dropout (high score) and learners very likely to stay in the course (low score). The weights are optimized and generalized to new data points using regularization.

- We use Random Forest [23], which tries to build different decision trees and then aggregates their votes to reach better performance. A decision is a sequence of binary decision that lead to a classification (dropout or active). By building several such decision trees we create different ways of discriminating samples which leads to a good performance when the prediction are averaged together. Here the number of trees built as well as the granularity of the decision are optimized.

- We also train a Nearest-neighbors [23] model which uses "close" examples in order to decide which category should be predicted for a new example. We optimize the number of neighbors used when computing the average.

The above optimization occurs when using cross-validation on the training course, which means for each algorithm we choose the above parameters that perform best on average when we train on 80% of the data and test on the remaining 20%. We choose the best model based on this cross-validation performance and use it to produce likelihood estimates.

It also should be noted that although the selected model can be applied to a different course, the prediction accuracy may decrease. Therefore, in real world scenarios, it is better to retrain the model with the data from the same course whenever the data is available.

4 VISUALIZATION DESIGN

In this section, we first introduce four design guidelines that have been summarized from the pilot interview. Subsequently, we describe the clustering method that we employ to cluster the learner groups and identify outliers. We then give a detailed description on the visual encoding for each view. Finally, we illustrate the interactions among different views.

4.1 Design Guidelines

We work closely with two domain experts on a monthly basis. A wide range of visual encoding decisions are considered throughout those discussions. Visual designs are formulated and decided after three rounds of opinion exchange and trials on sample data sets. We also identify four design guidelines from this iterative process.

G1. The visual designs should be easily understood by users with different backgrounds. Based on the previous work [41] and pilot interviews with domain experts, we realize that normally instructors do not have much knowledge in machine learning and statistical analysis. So they prefer intuitive visual designs that can be easily understood and better facilitate their exploration of the system. Therefore, the designs should have either familiar visual elements or metaphors that can be understood effortlessly. Our glyph design along the timeline is modified based on a common pie chart and the flow view has clear metaphors from daily life, which allow users to quickly understand the design and willingly use the system.

G2. The system should provide multiple layers to present information at different levels of granularity. Given that we have different levels of clustering for learner groups, multiple layers are required to show the information for each level. Although the layouts for those layers can be diverse, some general rules should be maintained, such as keeping the encoding consistent across all layers. Besides, it would be better if we could keep at least some hints about the previous layer after switching to another view. For example, when users perform individual analysis, the subgroup's timeline still remains on the screen for reference. In this way, users can compare the individual activity with that of the general learners from the same subgroup without constantly switching back and forth.

G3. The degree of user intervention should be balanced with automatic clustering methods. Although clustering methods can group learners without any user intervention, our end users have domain knowledge about specific courses and may intend to explore a particular learner group. Therefore, a certain level of user intervention should be provided for them to perform filtering on the data for a specific group. Moreover, given that most end users do not have a clear idea of learner group clustering especially at the beginning of the exploration, our system automatically recommends some group clusters. After investigating these recommended clusters, users can further merge and split the specified subgroups.

G4. System interactions should be familiar to end users and immediate feedback should be provided. As mentioned in the previous paragraph, user intervention is necessary and hence various interaction approaches should be implemented in the system. Due to the fact that people usually tend to explore the interactivity of a system by trials, more familiar operations are required for end users to perform interactions more fluently, such as brushing and filtering data attributes, dragging and dropping for learner groups, and lassoing circles for selecting a user-defined cluster. After performing those interactions, immediate feedback should be provided, such as highlights and fading, which could be much beneficial for users who are not familiar with such kind of systems.

4.2 Clustering Method

As mentioned, one of the most important analytical tasks for our end users is to understand the learning behaviors of a certain learner group. Although the domain knowledge of users may guide them to focus on a specific learner group, locating to a meaningful group by manually filtering requires much effort. Therefore, the automatic clustering method is desired. To detect learner clusters and outliers, we calculate a feature vector for each learner, and then perform dimension reduction to allow real-time clustering. A density-based clustering algorithm is then used to locate the learner groups and outliers. The steps are described as follows:

First, a feature vector for each learner has to be defined to represent the learning behavior of the learner. Given that the feature used in the predictive model is designed for the same purpose, we simply adopt it in the clustering method. A feature set commonly used for dropout prediction [20] is applied in our case study. The detailed definition of this feature vector is shown in Table 2.

Second, we use Multidimensional Scaling (MDS), which is one of the most commonly used algorithms for dimension reduction to reduce the original high dimensions into a 2D space. To measure the similarity between two features, we test Euclidean distance, Cosine distance and Canberra distance on a real dataset, and finally select Cosine distance since it is more sensitive to outliers.

Using the 2D feature, the clustering result can be updated in real-time to allow users to select a learner group interactively and further analyze the clustered subgroups. We can further accelerate the clustering algorithm by reducing the feature set to a 1D feature. However, the 2D feature adopted by our method can already support real-time clustering. Moreover, the 2D feature will be directly used to draw a scatterplot so that users can observe the relationships among different subgroups.

We then apply DBSCAN to cluster the learners into groups. DBSCAN presents two benefits in our scenario. First, DBSCAN is less time consuming so that we can update the clustering result in real-time. Second, it can detect both outliers and clusters without requiring a predefined number of clusters. To ensure that the detected clusters are consistent with the visual similarity in the scatterplot, we simply use the Euclidean distance in this step.

4.3 Visual Encodings

Based on the abovementioned design guidelines and the analytic tasks that are identified by the domain experts, we design a user-oriented interface for DropoutSeer. An overview of DropoutSeer is illustrated in Fig. 1, in which the cluster view on the left presenting the learner groups, the timeline view showing weekly activities and performances of different learner groups, the flow view in the middle illustrating the forum activities corresponding to each group, and the dashboard view on the right presenting various filtering and general information.

4.3.1 Cluster View

As shown in Fig. 1(a), the cluster view shows the clustering results of learner groups (T2). To identify different learner groups, we first classify learners based on their dropout and prediction results so that there are mainly four general categories: active-predicted, active-missed, dropped-predicted, dropped-missed. The general distribution of the four categories is illustrated at the top of the dashboard view (T1). Afterward, the cluster view presents the clustering results of each general category based on the clustering method that is described in Section 4.2. Given that the layout of cluster view is based on the 2D feature of the learner and MDS algorithm, we encode each learner as a dot. All dots that are related to the same cluster are highlighted whenever one of the dots is hovered on or when the cluster is selected in other views.

4.3.2 Timeline View

The timeline view aims to examine those factors that affect different learner groups (T2) and the learning temporal patterns that are demonstrated by these gruops (T3). Based on the clustering results, we separate different subgroups along the vertical axis and encode the temporal information along the horizontal axis. As

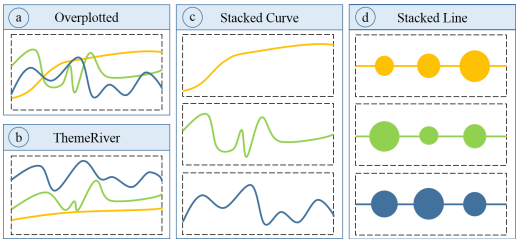

Figure 3: Layout alternatives for multivariate time-series. (a) the overplotted curve. (b) the Themeriver design. (c) the stacked curve and (d) the layout used in DropoutSeer.

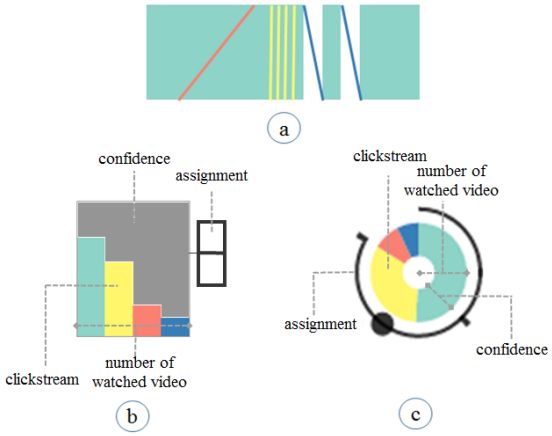

Figure 4: Alternatives for the glyph design. (a) the video-based glyph design shows the detailed clickstream records for each video. (b) and (c) are the bar chart based design and the donut chart based design respectively used in DropoutSeer.

shown in Fig. 1(b), each row represents a subgroup of learners' actions throughout the whole course period. To the right end of each row, a horizontal bar chart is drawn to encode the number of learners in that subgroup. A glyph is designed and aligned along the course timeline on a weekly basis to show the averaged clickstream and assignment records. At the initial stage of prototyping, we presented learner activities by day. However, we found that the majority of the learners were only active in one or two days. Moreover, given that most MOOCs release course videos and assignments on a weekly basis, separating different days in the same week would be unfruitful because one could not determine the day when highly valuable information would be provided. Upon these considerations and tests on the real dataset, we finally decided to illustrate the weekly video watching behaviors and assignment performance.

As shown in Fig. 4(c), the glyph is designed to show the learning pattern for each week. Specifically, the outer radius of the donut chart indicates the average number of videos watched by subgroup of learners and the inner radius shows the corresponding standard deviation of the same subgroup. Usually, subgroups with more active learners are more important for the analysis. Given that a glyph with small inner radius and large outer radius means the corresponding subgroup contains more active learners, therefore, the width of the donut chart can be interpreted as the confidence of the subgroup. The arc around the donut chart shows the average percentage and its standard deviation of achieved scores in each week's assignments. A complete circle denotes a full mark while a small-angle arc indicates the average poor performance for the subgroup in that week. The small circle on the arc represents the average percentage and the two short lines denote the standard deviation. Besides, the donut chart encodes the percentage of different click actions (i.e., play, pause, rewatch and skip). Given that the horizontal line is the course timeline, each column corresponds to a consecutive week. This donut chart based glyph is considered visually attractive by our end users, however, it might be difficult to compare sector sizes of two glyphs. Given that the differences between bar charts can be more accurately observed, we have also designed a bar chart based glyph (Fig. 4(c)). Similarly, the bar chart encodes the number of different click actions, and a simplified box plot shows the distribution of the assignment scores. The width of the whole bar chart indicates the average number of watched videos. Different from the donut chart, the bar chart uses the background color to encode the confidence of the subgroup. With the bar chart based glyph, users can compare different learner groups and explore the temporal data by investigating the timeline view either horizontally or vertically. The system uses the donut chart based design as default and users can switch between these two designs. Moreover, when a single user is selected in the flow view, an individual layer will be expanded for a more detailed analysis (T5). The visual encoding of the individual layer is similar with the design of a subgroup of learners, and the only difference lies in the fact that the attributes do not need to be averaged over a group of learners.

With regard to the scalability problem, the encoding components such as color and sizecan be scaled favorably without losing information. One concern raised by users is that the number of meaningful subgroups may greatly vary among different courses or general groups, whereas only a limited number of the subgroups can be shown simultaneously on the screen (e.g., four subgroups are shown in our case study). To address this issue, we have tuned our clustering method so that by default the number of clusters tends to be within a effective range. For example, a clustering result with only 100 small clusters may not be effective because it is too difficult for user to identify important subgroups. We have also ranked the subgroups based on the number of learners so that larger subgroups will be shown at the top of each general category. Users can merge and split the subgroups of interest, and reverse the rankings to observe outliers. Also, when users filter along various attributes in the dashboard, the cluster view and timeline view will be updated in real case accordingly. Finally, another design concern is how to encode the assignment behavior in one week. We choose to encode the relative percentage for correct assignments instead of the absolute scores, which is mainly because the total scores for the assignment can be different from week to week and some learners who join the course later tend to finish several assignments at a time. Thus, the absolute scores may not convey as much useful information as the accuracy of assignments calculated relatively. This choice has also been confirmed with the domain experts in our interviews.

Discussions on alternative designs.

Several alternative designs were taken into consideration before we decided to use this pie-chart based glyph on a straight timeline. First, several layout choices have been considered to display the timeline of different subgroups. As shown in Fig. 3, we can either plot different subgroups using the same timeline (Fig. 3(a) and (b)) or stack tiemlines over each other (Fig. 3(a) and (b)). When overplotting each subgroup on the same timeline (Fig. 3(a)), the view can easily becoming cluttered, not to mention further embedding glyphs on each curve. ThemeRiver design [14], or a curved timeline has also been considered. However, in this design, the statistics for each week is a discrete number, and it does not have consistent meaning similar to that in VisMOOC [30], which corresponds to the video timeline. We have sought the opinions of domain experts, and although they consider ThemeRiver/curved design to be more visually appealing, it still could not solve the comparison problem across different subgroups. Finally, we choose the simple stacked line for the reason that it can save more space to show more subgroups at the same time, and leave all other attributes to the glyph design.

For the glyph design, we initially presented the clickstream

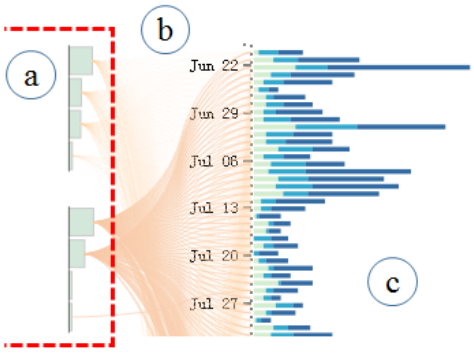

Figure 5: Illustration of the flow view design. (a) shows part of the timeline view and (c) shows the stacked bar chart of forum posts. (b) is the flow connects (a) and (c).

records for each video that the learners have watched by adding an extra video timeline in the glyph. As shown in Fig. 4(a), the horizontal axis of the glyph represents the video timeline, the green rectangles represent the video segments that have been watched, and the rectangles left blank represent the skipped segments. One clickstream record corresponds to a line that is plotted at the same timestamp. After testing on a real dataset, we find that this design is filled with too many details, which make the analysis and comparison of different videos, learner groups and weeks a very tedious process. The abundance of details will also hinder users from interpreting and generalizing the findings. Therefore, we only encode several integrated attributes in the glyph.

4.3.3 Flow View

The goal of the flow view is to allow users to correlate forum activities with the subgroup information(T4). As shown in Fig. 5(c), the vertical timeline on the right corresponds to the course timeline and learners from different groups contribute to the stacked bar charts whenever they release a post. The stacked bar chart shows the number of total posts each day. The three different types of posts (i.e., questions, discussions and others) are distinguished by color. Each flow in Fig. 5(b) denotes the posts from a subgroup to the timestamp indicated by the end point (Fig. 5(a)). To classify the post records into the three types, we employ the idea in an existing work [27] which classifies the posts for dropout prediction. To be more specific, we first extract lists of seed words for each type based on the course syllabus and then apply SeededLDA [18] to identify the topics of the posts. Flows are drawn using bezier curve and the control points are carefully selected so that the flows from the same subgroups are bundled together at the start point. Besides, the flows from other groups fade away when a learner group is selected.

4.4 Interactive Explorations

As mentioned in the design guidelines, interactivity is an important component in the DropoutSeer system. Guided by previous works [28, 40], we have considered a wide range of interactions and selected four of them based on our design requirements. These interactions facilitate users to explore the data freely at multiple levels and from different aspects.

Filtering. The dashboard view on the right allows users to filter along all attributes used to cluster learners. The axes are lined up by default according to their importance in current predictive model. However, users can still adjust the order of the axes so that all attributes of interest will be shown at the top of the dashboard. (G2)

Highlights. Selecting and highlighting often appear simultaneously to provide users with immediate feedback (G4). For example, when a subgroup is selected in the timeline view, the corresponding flows from that subgroup will be highlighted. In this way, users can easily align the two timelines even though they are not positioned in the same direction. Similarly, when the mouse hovers on a single post in the content view, the forum flow of this learner will be highlighted in the flow view.

Elaborate. When users click on a single day on the vertical timeline, the corresponding learner flows will be highlighted and all posts in that day will be shown at the bottom of the dashboard view. The background color of the post content is consistent with its corresponding type of post in the flow view (G2).

Reconfigure. By default, the subgroups are ranked by their sizes. We allow users to merge or split the subgroups. If a user find that two subgroups are similar and can be interpreted as one, the two groups could be merged by dragging and dropping, which is quite straightforward from the users perspective. similarly, the users can also split the group into two parts by double clicking on that group, or reverse the default ranking. (G3 and G4)

Table 2: The basic feature and the extended feature for predictive model and the corresponding feature importance calculated by the predictive model.

ID	Definition	Importance
f_{b1}	# of clickstream	0.20
f_{b2}	# of watched video	0.11
f_{b3}	# of active days	0.37
f_{b4}	# of play records	0.13
f_{b5}	# of pause records	0.40
f_{b6}	# of rewatch records	0.20
f_{b7}	# of skip records	0.01
f_{b8}	# of ratechange records	0.01
f_{b9}	# of posts	0.49
f_{b10}	marks of assignment	1.00

5 Case Studies

To evaluate the effectiveness and usefulness of DropoutSeer, we conduct case studies in collaboration with the instructors who offered courses on Coursera and edX, as well as researchers in the field of predictive modeling. We deploy the back-end part of the system on our server with a 2.7GHz Intel Core i7 CPU, 8GB memory PC, and the instructors could get access to the system through the web browser. In our case studies, we use the model that is trained on the data of the first month to predict the dropout rates of week five. The feature set we used for prediction is shown in Table 2. It is a commonly used feature set for dropout prediction [20]. Note that the system is flexible when the definition of dropout changes. For example, if we want to predict the dropout rates of week four based on the data of the first week, we only need to modify the corresponding settings in our data model. The data model will then retrain the predictive model based on the new training set and update the predictive results. The visualization system can directly visualize the updated data without any design modification.

Before the data can be explored in the system, it needs to be processed by the data manager and the data model. For the JAVA course used in our case study, the whole data cleaning and preprocessing take about 2.5 hours. To be more specific, the data cleaning process and training the predictive model take about 0.7 and 1.5 hours respectively, while the clustering algorithm and topic extraction algorithm only take a few minutes. Although the computational complexity is high, in practical scenarios, a daily update of the results is sufficient

Figure 6: The screenshot of DropoutSeer system for the *NCH* course. The distribution of each click type are clearly different from the *JAVA* course.

to provide users with up-to-date information. Therefore, the system can still be used for real-time analysis.

When the users explore the system, several patterns have been detected and the underlying insights are explained by the end users. We classify the major findings into three different categories and describe as the following.

In general, Fig. 1 and 6 show that the active and dropped learners have significantly different prediction accuracies. For both the *JAVA* and *NCH* course, the prediction accuracy of dropped learners is higher than that of the active learners. More specifically, the accuracy in the *JAVA* course is above 90% for dropped learners and approximately 50% for active learners. The *NCH* course shows similar patterns while the accuracy for active learners is even lower. The users also observe a clear positive correlation between the number of watched videos and learners performance, which is in accordance with the common sense that those learners who watch more videos tend to perform better in their assignments. When the users further check the clustering results, they find that the most inactive learners (e.g., drop at beginning of the course) form the largest subgroup in the dropped-predicted category. For other categories, we can see usually there are two to four subgroups and several outliers. For example, as marked in Fig. 6, there are four learners who participated in the course only on the prediction week. Therefore, They are misclassified as dropped learners.

When our users compare the *JAVA* course with the *NCH* course, a distinct difference is that in *JAVA* course, many learners who dropped from the fifth week came back and maintained active in the sixth week. By contrast, most learners who dropped at the fifth week in the *NCH* course did not come back (i.e., after week five, the size of the glyphs in the bottom part of Fig. 1 are larger than that in 6). Such difference may be attributed to the fact that all lecture videos are released to the learners before the end of the fifth week in the *NCH* course. In the following weeks, the students are only provided with some review tutorials and a final examination. However, for the *JAVA* course, another half of the lectures videos were going to be released after the fifth week, which motivates many dropped learners to return to the course. Therefore, whether a learner will come back after a whole week's silence is largely depends on the course itself. In real-time prediction, although some learners will be labeled as dropout at a certain time stamp, there is still some probability that they will return. To further confirm this finding, the users select those learners from *NCH* course who had no action in the second week. As marked in Fig. 7(a), many of these selected learners come back in the third week.

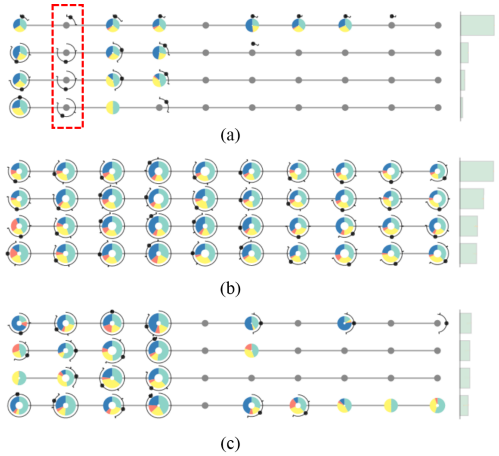

Figure 7: Patterns found in the timeline view. (a) learners have no action for a whole week may still come back if new course materials have been released. Different from users expectation, (b) learners who have a lot of skip actions tend to stay in the course rather than (c) drop the course.

5.1 Feature Importance of Predictive Model

The dashboard view (i.e., Fig. 1 and Fig. 6) shows that the records of the last week's assignment just before the prediction (i.e., the feature "problem4") are the most effective among all the features in the model. When we filter out those learners whose last week' s assignment behavior is lower than a threshold, all the other learners left will be those who keep active in the following week. During the exploration of the system, one user selects the learners who correctly answered more than 2/3 of all the fourth week's assignments and finds that all the learners in this group are correctly predicted and they remain active till the end of the course. It is noteworthy that there is no active-missed learners and no dropped learners in this group, which means the assignment behavior right before the timestamp for predicting dropout is extraordinarily important.

5.2 Abnormal Learner Subgroup

In contrast to the expectations of the users, who suggest that skip action means that the learners are not interested in the video content,

most of those learners who consistently skip lots of video segments have successfully finished their courses and received high grades at the end. For each of the first four weeks, the users filter the learners in order to find all those learners whose forward seeking activity is higher than 20 times per week. From Fig. 7(b) and (c), it is intuitive that most this kind of learners belong to the active categories and demonstrate a relatively favorable performance. One possible reason is that these learners have already had plenty prior knowledge of Java and therefore they only searched for useful materials in the videos. The users further validate this hypothesis from the forum posts. Some learners introduced their learning motivation as to update their knowledge of Java. Another possible reason is also pointed out by instructors when they explore our system that it may not be the first time for these learners to watch these videos. The learners also demonstrate many skip actions when reviewing their course materials for the final examination. Hardworking learners often understand their course materials thoroughly and only demonstrate forward-seeking behaviors to arouse their memories.

5.3 Long Tail in Dropout Distribution

When the users further explore the data, they find that the distribution of some attributes (i.e., rewatch, active days and number of posts) are similar with it for the skip action. Specifically, the users cannot easily distinguish active learners from dropped ones when the attribute has a small value, but they find only active learners remain when the value is larger than a threshold. Our collaborators from the machine learning field suggests that this finding can be helpful in prediction.

6 Expert Interview with Domain Experts

We also perform in-depth interviews with five experts to evaluate the usability of our visualization system. Among them, two course instructors (CIs) and one education expert (EE) have worked with us since the initial stage and have prior experience of using our previous systems for MOOC-related analysis. These are also two data mining researchers (DRs) who have worked on MOOC data for predicting dropouts in the 2015 KDD Cup [16], and both of them have never seen our system before.

Procedure. Each interview lasted for 60 minutes. During the interviews, we briefly introduced our project and gave the participants a tutorial on how to use the system. Afterward, the interviewees were requested to explore the system by themselves. Finally, we gathered their feedback on the usability, visual design and interactions of the system as well as solicited suggestions for potential improvements.

Overall system usability. Generally speaking, all the participants were satisfied with our visualization system and regarded it as intuitive to understand and easy to use. The first group of experts (two CIs and one EE) appreciated DropoutSeer as they have longed to see the joint analysis of video watching behavior, forum activities, and assignment results. One CI commented that "we have finally joined the pieces together". The second group of experts (DRs) who had no experience in using a visual analytics tool before were quite excited to see how learners activity could be visualized to help in dropout prediction. These researchers asked some questions about visual designs and development process.

Visual design and interactions. The MDS view was appreciated by both CIs and DRs. CI1 commented, "Clusters of different types of learners are clearer than the previous edition. Now it [the system] gives me a clear view of the subgroups." CI2 also regarded DropoutSeer useful and commented that "It helps me to see the general distribution of subgroups quickly and I could find some outliers from this view." He further selected those outliers and explore their detailed timeline view. He was satisfied with the capability of the system to achieve individual-level visual representation. The pie-chart-based glyph was quickly accepted and DR1 commented that "It [The glyph] is easy to remember once you understand it". EE particularly praised the flow view and regarded it as a natural illustration of linking learners' post threads with grouping information. The content view which linked with the flow view and timeline view was appreciated by CI2, who commented "I found the original posts quite useful. The classification of posts helped me identify valuable information. When I want to explore a post regarding learners' questions, I can now quickly locate it in the post threads instead of checking them one by one." Moreover, the experts appreciated the multiple interactions the system provided. CI1 commented that, "The filtering function allows me to explore the group of students in which I interested. The alignment of different axes based on importance also gives me some hints as to which features to filter." DR2 enjoyed the merge and split functions as he used drag and drop frequently when exploring different subgroups. Besides, both the course instructors and the education expert focused on learners who are near the dropout margin while data mining researchers paid more attention to the learner groups with wrong prediction results.

Limitations and suggestions. In the post-study interviews, the experts mentioned the limitations and provided many valuable suggestions. CI1 pointed out that self-paced learning would be the trend in MOOC, but the two courses in our work were still conservative ones with course materials being updated on a weekly basis. Also he suggested that short-period courses would be another trend and thus it would become harder for us to gather enough information to do prediction. "When dealing with ongoing courses, I was wondering if the system could provide on-the-fly analysis and prediction. The learners with a high potential of dropping out in the next week must be identified, so that I could make some adjustments or offer materials to those learners who fail to understand a specific knowledge key point." CI2 also commented, "Other than the dropout students, I am also interested in those students who participate actively in both forums and assignments yet still could not achieve good scores. Something must be done to help these students learn."

7 Conclusion and Future Work

In this paper, we propose a visualization system named DropoutSeer to help both instructors and machine learning researchers to analyze the relationship between the online learning activity and dropout behavior of learners. We collaborated with end users of this system to extract the analytical tasks and the design rationale as well as to build the analytical system. The case studies and the feedback from the experts confirmed the usefulness and effectiveness of the system.

Nevertheless, there are still some room for improvement. In the future, we plan to extend DropoutSeer in the following two directions. First, MOOCs are still in rapid development. The structure of the courses and the metric of evaluating learners continue to evolve over time. Therefore, we plan to design a more flexible analytical framework which can address this issue. For example, we can build an analytical model to detect the learning pace of learners or any periodic patterns in their learning activities in order for the time scale in the visualization system to be adjusted adaptively. Second, given that some instructors and education experts only interested in the analysis of certain types of learners, instead of predicting the general dropout behavior, we also wish to devise highly targeted visualization tools that can help to build a classification model for recognizing specific types of learners.

Acknowledgements

The authors would like to thank Prof. Raymond Chi-Wing Wong, Prof. Ting Chuen Pong and Mr. Tony W K Fung in HKUST for participating this project as domain experts, and the anonymous reviewers for their valuable comments. This work is supported by the Innovation Technology Fund of Hong Kong under Grant No. ITS/306/15FP.

REFERENCES

[1] W. Aigner, S. Miksch, H. Schumann, and C. Tominski. *Visualization of time-oriented data*. Springer Science & Business Media, 2011.

[2] B. Amnueypornsakul, S. Bhat, and P. Chinprutthiwong. Predicting attrition along the way: The uiuc model. In *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs*, pages 55–59, 2014.

[3] B. Bach, P. Dragicevic, D. Archambault, C. Hurter, and S. Carpendale. A review of temporal data visualizations based on space-time cube operations. In *Eurographics conference on visualization*, 2014.

[4] S. Boyer and K. Veeramachaneni. Transfer learning for predictive models in massive open online courses. In *Artificial Intelligence in Education*, pages 54–63. Springer, 2015.

[5] R. Brown, C. F. Lynch, M. Eagle, J. Albert, T. Barnes, R. Baker, Y. Bergner, and D. McNamara. Good communities and bad communities: Does membership affect performance? In *Proceedings of the 8th International Conference on Educational Data Mining*, 2015.

[6] J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. S. Ebert, and T. Ertl. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pages 143–152. IEEE, 2012.

[7] H. Chen, H. Atabakhsh, C. Tseng, B. Marshall, S. Kaza, S. Eggers, H. Gowda, A. Shah, T. Petersen, and C. Violette. Visualization in law enforcement. In *CHI'05 extended abstracts on Human factors in computing systems*, pages 1268–1271. ACM, 2005.

[8] Q. Chen, Y. Chen, D. Liu, C. Shi, Y. Wu, and H. Qu. Peakvizor: Visual analytics of peaks in video clickstreams from massive open online courses. 2015.

[9] C. Coffrin, L. Corrin, P. de Barba, and G. Kennedy. Visualizing patterns of student engagement and performance in moocs. In *Proceedings of the fourth international conference on learning analytics and knowledge*, pages 83–92. ACM, 2014.

[10] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. J. Gao, H. Qu, and X. Tong. Textflow: Towards better understanding of evolving topics in text. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2412–2421, 2011.

[11] F. Dernoncourt, C. Taylor, U.-M. OReilly, K. Veeramachaneni, S. Wu, C. Do, and S. Halawa. Moocviz: A large scale, open access, collaborative, data analytics platform for moocs. In *NIPS Workshop on Data-Driven Education, Lake Tahoe, Nevada. Retrieved from http://groups. csail. mit. edu/EVO-DesignOpt/groupWebSite/uploads/Site/MoocViz. pdf*, 2013.

[12] M. Dörk, D. Gruen, C. Williamson, and S. Carpendale. A visual backchannel for large-scale events. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1129–1138, 2010.

[13] P. Dragicevic and S. Huot. Spiraclock: a continuous and non-intrusive display for upcoming events. In *CHI'02 extended abstracts on Human factors in computing systems*, pages 604–605. ACM, 2002.

[14] S. Havre, B. Hetzler, and L. Nowell. Themeriver: Visualizing theme changes over time. In *Information Visualization, 2000. InfoVis 2000. IEEE Symposium on*, pages 115–123. IEEE, 2000.

[15] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. Themeriver: Visualizing thematic changes in large document collections. *Visualization and Computer Graphics, IEEE Transactions on*, 8(1):9–20, 2002.

[16] http://kddcup2015.com/information.html. KDD2015, year = 2015, url = http://kddcup2015.com/information.html, urldate = 2010-09-30.

[17] J. Huang, C. Piech, A. Nguyen, and L. Guibas. Syntactic and functional variability of a million code submissions in a machine learning mooc. In *AIED 2013 Workshops Proceedings Volume*, page 25. Citeseer, 2013.

[18] J. Jagarlamudi, H. Daumé III, and R. Udupa. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213. Association for Computational Linguistics, 2012.

[19] J. Kim, P. J. Guo, D. T. Seaton, P. Mitros, K. Z. Gajos, and R. C. Miller. Understanding in-video dropouts and interaction peaks inonline lecture videos. In *Proceedings of the first ACM conference on Learning@ scale conference*, pages 31–40. ACM, 2014.

[20] M. Kloft, F. Stiehler, Z. Zheng, and N. Pinkwart. Predicting mooc dropout over weeks using machine learning methods. In *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs*, pages 60–65, 2014.

[21] S. Liu, W. Cui, Y. Wu, and M. Liu. A survey on information visualization: recent advances and challenges. *The Visual Computer*, 30(12):1373–1393, 2014.

[22] S. Liu, M. X. Zhou, S. Pan, Y. Song, W. Qian, W. Cai, and X. Lian. Tiara: Interactive, topic-based visual text summarization and analysis. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(2):25, 2012.

[23] D. Michie, D. J. Spiegelhalter, and C. C. Taylor. Machine learning, neural and statistical classification. 1994.

[24] W. Müller and H. Schumann. Visualization methods for time-dependent data-an overview. In *Simulation Conference, 2003. Proceedings of the 2003 Winter*, volume 1, pages 737–745. IEEE, 2003.

[25] D. F. Onah, J. Sinclair, and R. Boyatt. Dropout rates of massive open online courses: behavioural patterns. *EDULEARN14 Proceedings*, pages 5825–5834, 2014.

[26] C. Plaisant, B. Milash, A. Rose, S. Widoff, and B. Shneiderman. Lifelines: visualizing personal histories. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 221–227. ACM, 1996.

[27] A. Ramesh, D. Goldwasser, B. Huang, H. D. Iii, and L. Getoor. Understanding mooc discussion forums using seeded lda. 2014.

[28] K. Sedig and P. Parsons. Interaction design for complex cognitive activities with visual representations: A pattern-based approach. 2013.

[29] M. Sharkey and R. Sanders. A process for predicting mooc attrition. In *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs*, pages 50–54, 2014.

[30] C. Shi, S. Fu, Q. Chen, and H. Qu. Vismooc: Visualizing video clickstream data from massive open online courses. In *Visualization Symposium (PacificVis), 2015 IEEE Pacific*, pages 159–166. IEEE, 2015.

[31] L. Shi, F. Wei, S. Liu, L. Tan, X. Lian, and M. X. Zhou. Understanding text corpora with multiple facets. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 99–106. IEEE, 2010.

[32] T. Sinha, P. Jermann, N. Li, and P. Dillenbourg. Your click decides your fate: Inferring information processing and attrition behavior from mooc video clickstream interactions. *arXiv preprint arXiv:1407.7131*, 2014.

[33] T. Sinha, N. Li, P. Jermann, and P. Dillenbourg. Capturing" attrition intensifying" structural traits from didactic interaction sequences of mooc learners. *arXiv preprint arXiv:1409.5887*, 2014.

[34] C. Taylor, K. Veeramachaneni, and U.-M. O'Reilly. Likely to stop? predicting stopout in massive open online courses. *arXiv preprint arXiv:1408.3382*, 2014.

[35] J. Thomas and J. Kielman. Challenges for visual analytics. *Information Visualization*, 8(4):309–314, 2009.

[36] J. J. Van Wijk and E. R. Van Selow. Cluster and calendar based visualization of time series data. In *Information Visualization, 1999.(Info Vis' 99) Proceedings. 1999 IEEE Symposium on*, pages 4–9. IEEE, 1999.

[37] M. Weber, M. Alexa, and W. Müller. Visualizing time-series on spirals. In *Infovis*, page 7. IEEE, 2001.

[38] Y. Wu, N. Pitipornvivat, J. Zhao, S. Yang, G. Huang, and H. Qu. egoslider: Visual analysis of egocentric network evolution. *Visualization and Computer Graphics, IEEE Transactions on*, 22(1):260–269, 2016.

[39] P. Xu, Y. Wu, E. Wei, T.-Q. Peng, S. Liu, J. J. Zhu, and H. Qu. Visual analysis of topic competition on social media. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2012–2021, 2013.

[40] J. S. Yi, Y. ah Kang, J. Stasko, and J. Jacko. Toward a deeper understanding of the role of interaction in information visualization. *IEEE transactions on visualization and computer graphics*, 13(6):1224–1231, 2007.

[41] C. Zinn and O. Scheuer. Getting to know your student in distance learning contexts. In *Innovative Approaches for Learning and Knowledge Sharing*, pages 437–451. Springer, 2006.