Semester Project Report: MOOC Clickstream Visualization

Yumeng Hou
yumeng.hou@epfl.ch

1 Introduction

Massive Open Online Courses (MOOCs), which aims at unlimited participation and open access to education, have attracted considerable public attention in the last few years. EPFL is actively involved in this trend and strikes to be one of the pioneers. By the end of 2014, EPFL had released 24 online courses, with additional 26 in preparation. Among the released MOOCs, there are 5 main categories depending on the academic level: perparatory, propedeutic, Bachelor, Master, and extra curricular. All the courses are given in English or French with correspondent subtitles [1].

Although saying goes that MOOC is reshaping the higher education with wide and open access to education via the Internet, it suffers from limited face-to-face interactions between teachers and students. Lack of direct observation on participation consequently poses a big challenge for course providers to understand students' learning behaviors and improve their teaching accordingly. Fortunately, researchers could "observe" the learners' participation based on the large-scale data acquired via MOOC technologies. Clickstream data analysis is capable to provide insight into learners' behaviors and learning patterns, which are significant for instructors to revise course materials accordingly.

This project is dedicated to interpretative visualization of MOOC clickstream data. Guided by the collaborative nine-stage design study methodology framework [14], we began with a literature review of the state-of-art research on MOOC analysis to understand current practices and challenges, as presented in Section 2. Afterwards, we applied suitable statistical techniques and simple visualization tricks for explorative understanding on the collected MOOC clickstream data, which will be introduced in Section 3. We set a clear goal for the tasks and scheduled the milestones, as described in Section 4. Methodologies utilized to achieve data preprocessing and pattern exploration will be clarified in Section 5. Section 6 is to describe the 3 iterations of visualization designs, implementations and results. In Section 7, I will discuss on evaluations and future works.

2 Literature Review

2.1 Online Education Analysis

There have been a lot of work targeting to analyze online learning behavior, where statistics and basic visualizations have been widely used. The main goals of these studies can be classified into several categories:

- Student activities and patterns of participation [9]; Student's demographic information and its relationship with learning styles [19];
- Forum interactions, including social network visualization in online learning groups [11], community relationship in peer-to-peer systems [12] and patterns of time-varied forum activities;
- Student performance: including grades on assignments, quizzes and exams [15].

Within the field, some visualization tools have been developed to provide various visual representations and allow users to interact based on their specific goals. CourseVis [8] is a course management system which aims to help instructors become aware of social, behavioral and cognitive aspects of e-learners. It is contributive in presenting a three-dimensional scatterplot for web log data but weak in handling the scale of MOOCs. E-learning tracking [5] demonstrates a set of loosely coupled visualization tools that help to display and analyze student interactions with online courseware. These tools mainly focus on student access to course materials and the navigation path which a student follows throughout the course. SST [2] presents an interactive visualization design for temporal activity patterns, where the timeline spiral graph is combined with other two inter-linked supporting panels.

2.2 Clickstream Visual Analysis

One of the greatest advantages of MOOCs is that it provides us with a massive amount of behavioral data stream. Apart from basic page view and forum logs, MOOC platforms keep track of student interaction data at a higher level. For example, the various click actions that happen during the learning process. Researchers have been studying this kind of clickstream data for decades before MOOCs. While very little work has been done to interactively visualize the time series of massive open online courses and combine it with content-based analysis. CLAS [11] is a collaborative video annotation tool based on explicit user data by recording user clicks around points they are interested in. A series of MOOC analysis systems have been proposed to analyze in-video dropouts and interaction peaks [6], video production styles with student engagement [3], and student demographic differences in navigation behavior [4]. One example is VisMOOC [17]. It makes to present an expert review with rich interactions, and allows instructors and educators to do further analysis based on their domain knowledge. The limitation is that it only analyzes student leaning behaviors based on clickstream within the lecture videos.

2.3 Clickstream Visual Representations

Existing research on clickstream visualization covers diverse areas. Some tools target at online shopping click sequence [7], while others concentrate more on interaction of users with videos [17]. These visual representations focus more on the transition of page browsing and click sequence. For example, [16] utilizes horizontal stacked bar to visualize a sorted list of web sessions after aggregation, while [18] explores visual clusters of web clickstream data through an intuitive user interface. VisMOOC [17] is inspiring by applying parallel coordinates to present the click-behavior stream within one video, where two parallel axes lines between are drawn to describe the starting and ending positions of seek events (Figure 1). Intuitive color coding is adopt in this representation to demonstrate different directions of seek events which indicate different learning behaviors.

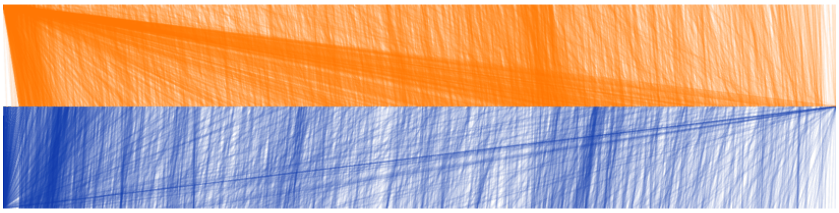

Figure 1: Parallel Coordinates of Video clickstream by VisMOOC [17].

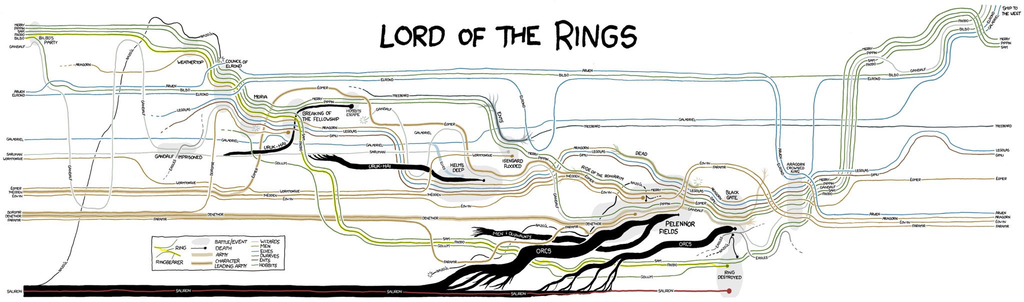

Figure 2: XKCD's Hand-drawn Storyline Illustration "Movie Narrative Charts" .

Some existing solutions make it possible for analysts to extract user behavior patterns and dig out clusters. But very few can be implemented to interpret the evolution of behavior patterns intuitively. On one hand, visualization of such large-scale data stream as MOOC clickstream is always challenging to overcome visual clutter or provide storytelling intuition. On the other hand, previous analytical works tend to focus more on tracking student behaviors while neglect the transitions among leaning patterns. Storyline is a good option to solve the design problem in storytelling. For example, in XKCD's [20] hand-drawn illustration "Movie Narrative

Charts" (Figure 2), each movie character's life-span was represented by a chronological line which converges and diverges with other characters' lines, depicting their interactions. Storyline is able to portray the temporal dynamics of social interactions by projecting the timeline of the interaction onto an axis. The challenge is to fit it into MOOC context where there are a huge amount of students which could be considered as "characters". Another alternative is Sankey diagram [13, 10]. Sankey diagram (Figure 3) is a type of flow diagram typically used to stress a visual emphasis on the transfers or flows within a system or between processes. In Sankey diagrams, arrows in different colors are demonstrating the transition status of different objects in time series and width of the arrows are shown proportionally to the flow quantity. Thus it is capable to illustrate evolutional flow while preserving information on transition and quantity.

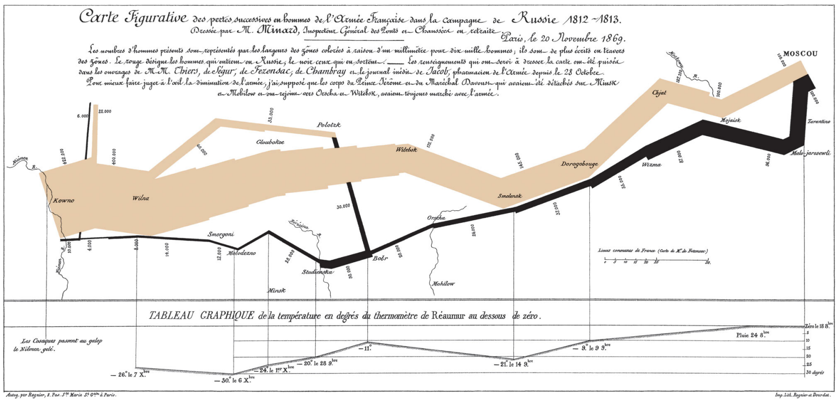
Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.
Kowno
Wilna
Smorgoni
Molodetschno
Minsk
Studienska
Bobr
Polotzk
Gloubokoe
Witebsk
Orscha
Mohilow
Smolensk
Dorogobouge
Wizma
Chjat
Mojaisk
MOSCOU
Tarantino
Malo-jarosewli
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

Figure 3: Sankey Diagram Example: Charles Minard's Map of Napoleon's Russian Campaign.

3 Problem Characterization

3.1 MOOC Data Description

There are four major types of data made available by major MOOC platforms:

- Registration data, which shows the amount and origin of course subscribers (students);
- Demographic data, that provides personal information such as age, gender, location and expertise;
- Engagement data, namely the student logs and clickstream. It provides information on which video a student has watched, how one navigates the videos, solves problems and interacts in the forum. This type of data is far more massive than those from other sources and is considered to be most contributive in our research context;
- Achievement data, that provides the grades obtained by one in assignments, problems, exams and finally for the course, as well as the certificates earned.

3.2 Dataset Description

In our case of this project, the dataset contains IDs (UserID), clickstream and behavioral information (UserEvents, Sessions) and achievement information (AchievementLevel, FinalGrade) of 17433 users. Figure 4 demonstrates the contents and structure in detail.

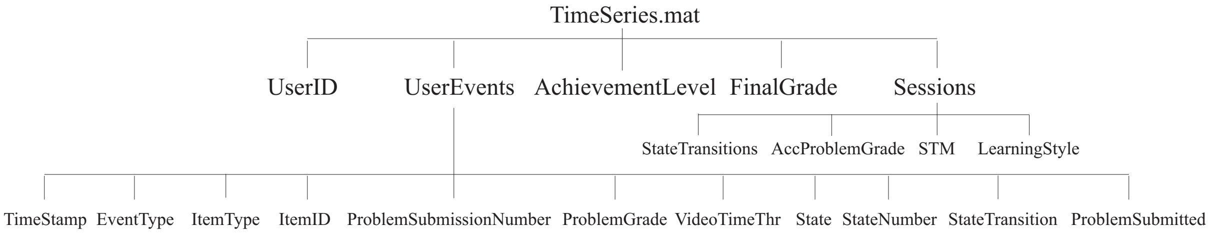

Figure 4: The Structure of our Dataset.

3.3 Markov Chains & Transition Matrix

To begin with, we segment the clickstream data into different sessions based on time interval, one week for each and 13 sessions in total. We extract Markov chains of the behaviors of each user within one session. The rule to extract Markov chains of the state machine is shown in Figure 5. Video.Load represents student behavior of watching videos, while whether the ItemID (of lecture video) is new or not means whether the student is learning new things or reviewing old materials. Problem.Check is the event related to problem submissions, namely the participation of exercises. Forum.X.View demonstrates the reading behaviors in the forums. While the rest state with 3 different sets of forum actions means active participation in the forums.

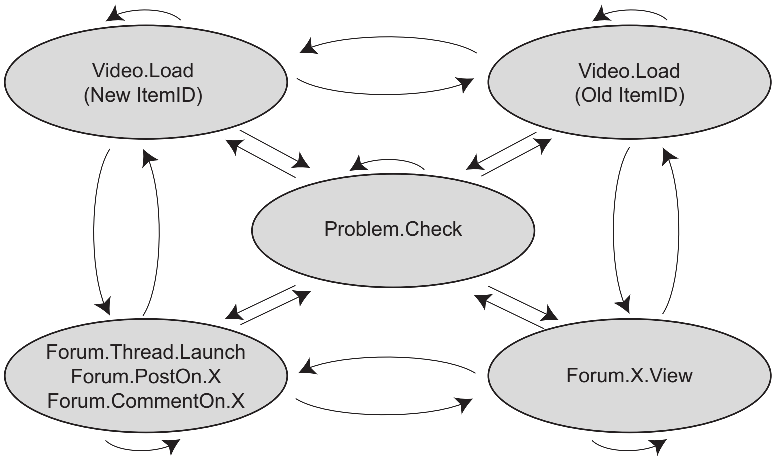

Figure 5: Rule to Extract Markov Chains

The Markov chains are significant for our goal of visualization as they are basis on which we conduct proper clustering method to calculate different student learning patterns. We conduct k-means clustering on all the extracted Markov chains and calculate based on the transitions. As a result, there are 6 distinct clusters (patterns). Based on the transition matrix of centroids, we manually interpret each pattern as follows:

- Pattern 1: “Idle”: Students are mostly inactive during the session since there is few transitions among states.
- Pattern 2: “Input”: Students focus more on learning new subjects since transitions are dominated by loops of Video.Load (New ItemID).
- Pattern 3: “Review”: Students focus more on reviewing old subjects as most transitions are going to Video.Load (Old ItemID).
- Pattern 4: “Exercise”: Students focus more on exercising since loops of or transitions are going to Problem.Check are overwhelming.
- Pattern 5: “Search”: Students focus more on reading forums as transitions are dominated by those going to Forum.X.View.
- Pattern 6: “Interact”: Students focus more on interacting in forums since transitions to the active Forum state are intensive.

Finally, based on the generated clusters and patterns, we output a new matrix containing pattern chains of all students. This matrix will be used as the input data for later computations and visualizations.

4 Tasks & Milestones

To finish such a semester project which lasts for 14 weeks, I personally set up 4 milestones aiming at different tasks along the process.

- Milestone 1: Data preprocessing and exploratory visualization

- Milestone 2: Basic interpretation with statistical methods
- Milestone 3: Visualization design
- Milestone 4: Visualization implementation
- Milestone 5: Evaluation and discussion

The following sections are to present the results along my milestones. Section 5 (Methodology) involves Milestone 1 & 2, Section 6 (Visualization) displays Milestone 3 & 4, and Section 7 (Discussion) is on Milestone 5.

5 Methodology

5.1 Programming Tools

5.1.1 Matlab R2015b

I use Matlab R2015b to achieve data preprocessing and transformation, as well as computations to generate student prototypes (which will be explained in Section 5.3.1). Matlab is also utilized for basic statistical analysis and exploratory visualization (e.g. distribution histogram and transition matrix).

5.1.2 d3.js

d3.js is a JavaScript library for visualizing data with HTML, SVG, and CSS. It is efficient in visual binding and interactive representations. I use d3 to ultimately implement my visualization designs, including parallel coordinates, storyline and Sankey diagram.

5.2 Outlier Removal

According to basic exploratory visualization, a huge amount of users are mostly idle without meaningful activities which result in a bias within the dataset and greatly influence the computation and visualization. Based on our observation, we remove the entries of student who score below 20 as outliers. The dataset consequently shrinks from 17433 entries to 661.

5.3 Additional Preparations for Visualization

5.3.1 Extraction of 50 Prototypes

I apply basic k-means clustering method to the pattern chains, with k = 50, on the clean dataset of 661 users. Thus, 50 student prototypes (centroids) are got. I use the data of student prototypes as alternative input for visualization to get a “representative visualization” and to compare with the original result.

5.3.2 Data Conversion

When visualizing data stream, especially for flow-like representations such as parallel coordinates and Sankey diagram, it’s always necessary to convert table/matrix-like data to a structure where data is organized in the form of {node, link, weight}. I use Matlab to traverse all the pattern chains, and calculate the weight of vector link from each source node to target node (node here refers to one pattern in one session). I output the data with information of nodes, links and weights to a new JSON file.

6 Visualization

The major challenge of visualization design in this project lies in to present the continuous transitions of learning patterns of such large amount of users intuitively and without visual clutter. To overcome this challenge, I went through three iterations of tryouts. While designing, I keep the simple principle and follow the Gestalt Law of continuity.

6.1 Line-based Visualization

Line chart is often used to visualize a trend in data over intervals of time, namely a time series. My idea starts with designing a variant of line chart to depict the evolutional information of learning patterns.

6.1.1 Parallel Coordinates

Following the idea of [17], previously shown in Figure 1, I designed a series of vertical parallel coordinates (Figure 6), to demonstrate the flow of pattern transitions between each pair of neighboring sessions. The result was not acceptable since as time goes on, it's increasingly harder to observe a clear path of pattern transitions among the students. Synchronizing approach is necessary as a basic to present with parallel coordinates for a large amount of contributors, but it's not easy to achieve.

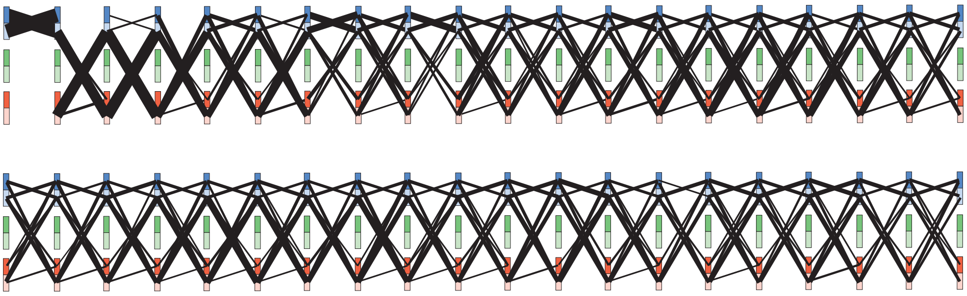

Figure 6: Parallel Coordinate Series Representation of Pattern Transitions.

6.1.2 Storyline

Under the inspiration from XKCD's [20] appealing hand-draft (Figure 2), I try to adapt storyline for MOOC visualization. I regard each user as a character and portray their life-span of MOOC with individual lines. I consider each session as a battle where characters intensively interact with MOOC (e.g. exercises) and casualty happen (e.g. some students drop the class). I assign different colors to the lines based on students' final grades (Fail: Grey; Pass: Light Green; Pass with good grades above 75: Dark Green). After generating the basic version of storyline (shown in Figure 7), I realize this kind of graph is less-effective in depicting too many characters, here 661 students. It makes to help indicate via color density, though not clearly, that there is a relation between grade distribution and learning patterns, that green and dark green lines are denser in the upper part representing forum participation and lecture reviewing. However, the graph turns out to too messy to figure out evolutional paths.

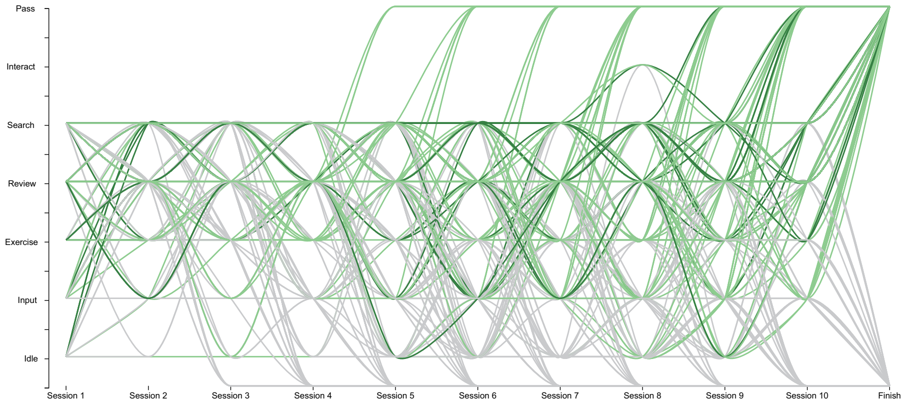

Figure 7: Storyline Representation of Pattern Transitions.

6.1.3 Storyline of Prototypes

The Storyline of all the users causes a heavy visual clutter and thus results to less effectiveness in clear representation. I change the input data with the 50 prototypes as described previously in Section 5.3.1. As shown in Figure 8, the representative storyline turns out to be much clearer. It is obviously indicated by dark green lines that there is a positive relation between good grades with forum reading and lecture reviewing behaviors.

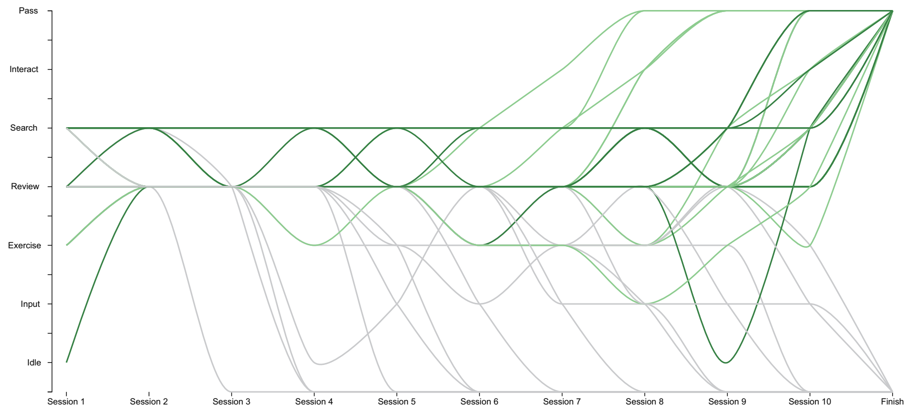

Figure 8: Storyline Representation of Pattern Transitions with 50 Prototypes.

6.2 Sankey Diagram

Sankey diagram is a promising in visualization approach to represent proportional transferring process of different patterns among sessions. In our case, it is expected to effectively represent evolutional information of when and how much proportion of students change from one learning pattern to another.

6.2.1 Visual Coding & Layout

As shown in Figure 9(d), I distribute a set of colors to the six patterns. I add two states "Pass" and "Fail" and assign them colors of green and grey respectively. The horizontal axis (Figure 9(a)) represents the sequential information: the 13 sessions and the final stage. The vertical axis (Figure 9(b)) is partitioned into bars by different patterns appearing in this session proportionally and in order. The band (Figure 9(c)) connecting one bar in the previous session to the current one represents a transition of patterns between sessions. Its color is dominated by the target pattern and its width represents the proportional volume of this transitional group.

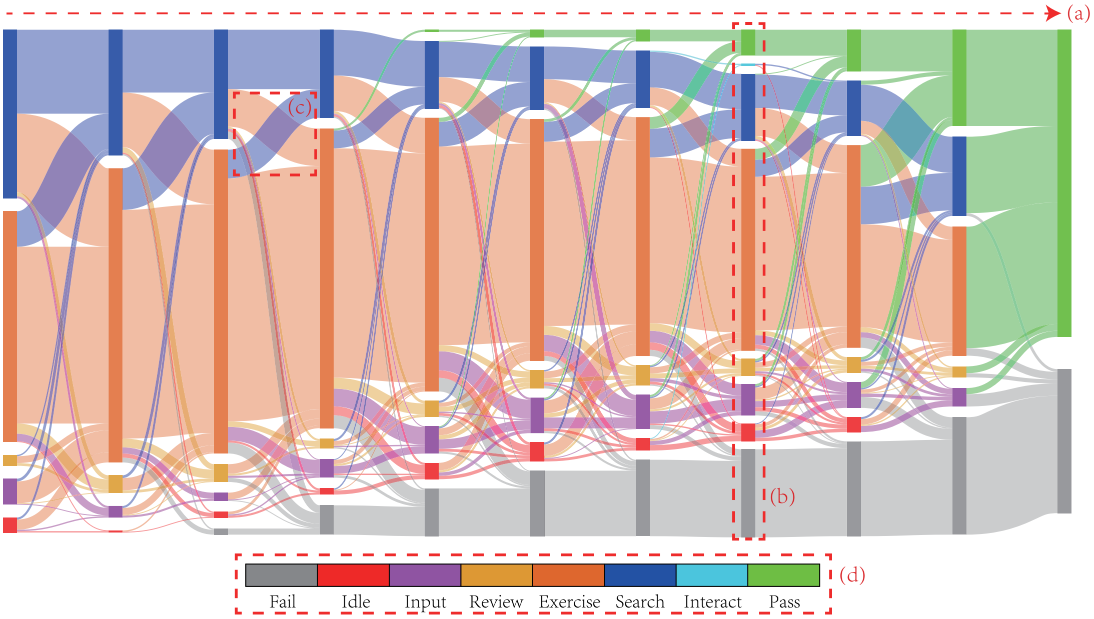

Figure 9: Sankey Diagram Representation of Pattern Transitions.

6.2.2 Sankey Diagram of Prototypes

The Sankey diagram above already presents an evolutional process of pattern transitions. This representation is significant, but with some disturbing branches. By changing the input data with the 50 prototypes, the visualization (Figure 10) becomes cleaner with more significant patterns on student's evolutional transition paths.

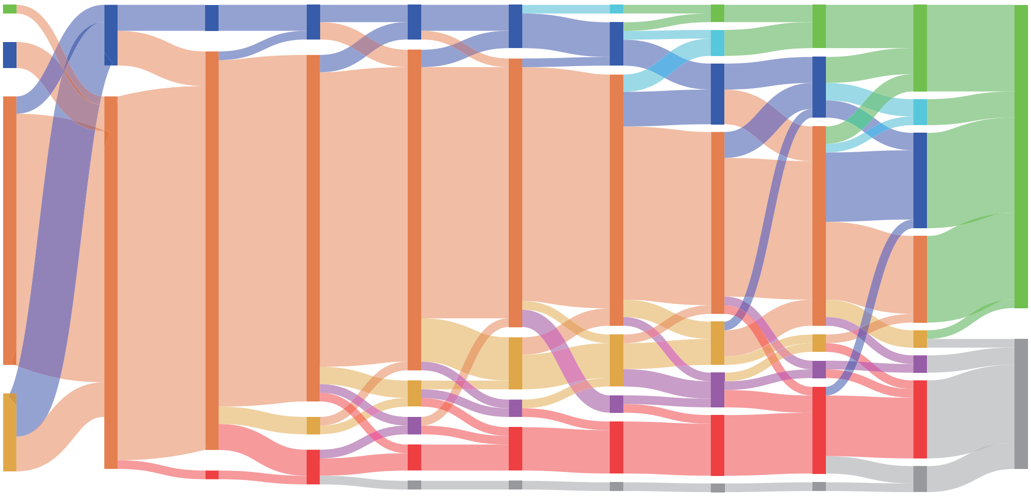

Figure 10: Sankey Diagram Representation of Pattern Transitions with 50 Prototypes.

Based on the Sankey results, a majority of students' learning behaviors are solving problems in the exercise to gain points, some of them even stop learning after reaching enough scores to pass the course. Students with a learning pattern of focusing on reading forums tend to stay active in forums during the whole learning process. And as the final exam coming near, forum participation gets more valued among students. To summarize, students who are inactive or only focusing on watching videos are more easy to fail the course, while solving problems and forum participations tend to lead students to success.

7 Discussion

In this project, we utilize MOOC clickstream data to generate Markov chains, based on which learning patterns are recognized with clustering methods. With the data of student learning patterns, I conduct three iterations of visualization design and implementation after a literature review of the state-of-art research. As a result, three visual representations were tried out: parallel coordinate series, storyline and Sankey diagram.

In terms of interpretation, Sankey diagram is most effective in demonstrating evolutional path of pattern transitions as well as indicating relationship between grades and student learning patterns. Storyline is good to give a simple impression of the relation between student grades and learning patterns. But it's not an efficient indicator for transitional evolutions. Parallel coordinate series is not a good solution since it's not so expressive and needs synchronization which is hard to achieve.

From the perspective of visual aesthetics, Sankey diagram is the best and storyline representation comes second, according to the feedbacks in my interviews. 5 out of 8 interviewees considered Sankey diagram to be intuitive and easy to understand, while 3 of them favored storyline. 6 out of 8 interviewees agreed that visualization of 50 prototypes were more expressive in evolutions while 2 persons worried that some information might be lost in this process.

For future work, on one hand, I want to optimize storyline representation. Storyline is a clear way to convey interesting stories in time series, but it doesn't work well with too many characters. For this case of MOOC, one alternative is to consider different learning patterns as characters, and specific sessions as battles. In this way, it is possible to dig out when and why users tend to change their learning strategies (when patterns die). On the other hand, I want to work further with the system development. The goal is to make the visualization more expressive and to allow analytics to conduct interactive visual explorations and analysis in deep with MOOC clickstream data.

References

[1] Epfl mooc report 2012-2014. *moocs.epfl.ch*, 2014.

[2] Diego Alonso Gómez Aguilar, Roberto Therón, and Francisco José García Peñalvo. Semantic spiral timelines used as support for e-learning. *J. ucs*, 15(7):1526–1545, 2009.

[3] Philip J Guo, Juho Kim, and Rob Rubin. How video production affects student engagement: An empirical study of mooc videos. In *Proceedings of the first ACM conference on Learning@ scale conference*, pages 41–50. ACM, 2014.

[4] Philip J Guo and Katharina Reinecke. Demographic differences in how students navigate through moocs. In *Proceedings of the first ACM conference on Learning@ scale conference*, pages 21–30. ACM, 2014.

[5] Judy Hardy, Mario Antonioletti, and S Bates. e-learner tracking: Tools for discovering learner behavior. In *The IASTED International Conference on Web-base Education*, pages 458–463, 2004.

[6] Juho Kim, Philip J Guo, Daniel T Seaton, Piotr Mitros, Krzysztof Z Gajos, and Robert 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.

[7] Juhnyoung Lee, Mark Podlaseck, Edith Schonberg, and Robert Hoch. Visualization and analysis of clickstream data of online stores for understanding web merchandising. In *Applications of Data Mining to Electronic Commerce*, pages 59–84. Springer, 2001.

[8] Riccardo Mazza and Vania Dimitrova. Coursevis: A graphical student monitoring tool for supporting instructors in web-based distance courses. *International Journal of Human-Computer Studies*, 65(2):125–139, 2007.

[9] Emmanuel N Ogor. Student academic performance monitoring and evaluation using data mining techniques. In *Electronics, Robotics and Automotive Mechanics Conference, 2007. CERMA 2007*, pages 354–359. IEEE, 2007.

[10] Patrick Riehmann, Manfred Hanfler, and Bernd Froehlich. Interactive sankey diagrams. In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*, pages 233–240. IEEE, 2005.

[11] Evan F Risko, Tom Foulsham, Shane Dawson, and Alan Kingstone. The collaborative lecture annotation system (clas): A new tool for distributed learning. *Learning Technologies, IEEE Transactions on*, 6(1):4–13, 2013.

[12] Cristóbal Romero and Sebastián Ventura. Educational data mining: a review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 40(6):601–618, 2010.

[13] Mario Schmidt. The sankey diagram in energy and material flow management. *Journal of industrial ecology*, 12(1):82–94, 2008.

[14] Michael Sedlmair, Miriah Meyer, and Tamara Munzner. Design study methodology: Reflections from the trenches and the stacks. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2431–2440, 2012.

[15] Ruimin Shen, Fan Yang, and Peng Han. Data analysis center based on e-learning platform. In *The Internet Challenge: Technology and Applications*, pages 19–28. Springer, 2002.

[16] Zeqian Shen, Jishang Wei, Neel Sundaresan, and Kwan-Liu Ma. Visual analysis of massive web session data. In *Large Data Analysis and Visualization (LDAV), 2012 IEEE Symposium on*, pages 65–72. IEEE, 2012.

[17] Conglei Shi, Siwei Fu, Qing Chen, and Huamin Qu. Vismooc: Visualizing video clickstream data from massive open online courses. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*, pages 277–278. IEEE, 2014.

[18] Jishang Wei, Zeqian Shen, Neel Sundaresan, and Kwan-Liu Ma. Visual cluster exploration of web clickstream data. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pages 3–12. IEEE, 2012.

[19] Fionán Peter Williams and Owen Conlan. Visualizing narrative structures and learning style information in personalized e-learning systems. In *Advanced Learning Technologies, 2007. ICALT 2007. Seventh IEEE International Conference on*, pages 872–876. IEEE, 2007.

[20] XKCD. Movie narrative charts. In *https://xkcd.com/657/*.