Topic-wise Classification of MOOC Discussions: A Visual Analytics Approach

Thushari Atapattu, Katrina Falkner, Hamid Tarmazdi
School of Computer Science
University of Adelaide
Adelaide, Australia
{firstname.lastname}@adelaide.edu.au

ABSTRACT

With a goal of better understanding the online discourse within the Massive Open Online Course (MOOC) context, this paper presents an open source visualisation dashboard developed to identify and classify emergent discussion topics (or themes). As an extension to the authors' previous work in identifying key topics from MOOC discussion contents, this work visualises lecture-related discussions as a graph of relationships between topics and threads. We demonstrate the visualisation using three popular MOOCs offered during 2013. This work facilitates the course staff to locate and navigate the most influential topic clusters as well as the discussions that require intervention by connecting the topics with the corresponding weekly lectures. Further, we demonstrate how our interactive visualisation can be used to explore correlation between discussion topics and other variables such as views, posts, votes, and instructor intervention.

Keywords

Visualisation, learning analytics, topic model, MOOC, online discourse, discussion forum.

1. INTRODUCTION

Within the educational context, visualisation of learning analytics, often known as 'visual analytics', provides insights for many end users including teachers, learners, researchers, educational platform developers, and institutions. According to Thomas and Cook [1], visual analytics focuses on analytical reasoning facilitated by interactive visualisation interfaces. In the educational context, visual analytics support teachers in identifying at-risk students, analysing students' engagement and performance of the course, social collaborations, and developing analytics on the students' online discourse. Visualisation dashboards also support self-evaluation for students in reflecting on their own learning process, setting goals and monitoring progress to achieve these goals.

Visual analytics are often useful in large to massive classrooms such as Massive Open Online Courses (MOOCs), facilitating the understanding of interesting patterns in large volume of students' data, which is challenging to observe using statistical analysis. Visualising the patterns of student engagement (e.g. lecture/forum view), behavior, social interactions and their relationship with final grade/performance has being a focus of many studies [2-4].

Even though the *system-generated* analytics on students' engagement and behavior are important to identify patterns that positively correlate with the successful learning outcomes or attrition, it is likely that these can generate some inconsistencies. For instance, a download of a lecture does not necessarily imply student engagement. Similarly, it is uncertain whether an up-vote of a forum post means the learner has an interest in the content or, alternatively, that they have problems associated with the topic discussed in the post. Therefore, the analysis of *learner-generated* online discourse (i.e. content) facilitates the interpretation of learners' cognitive processes as well as situating learner behavior in context. According to Mercer [5], the sociocultural perspective highlights "the possibility that educational success and failure may be explained by the quality of educational dialogue, rather than simply in terms of the capability of individual students or the skill of their teachers". This includes identification of individual's understanding of – and interest in – particular course content, and their level of expertise and activity in seeking assistance to rectify conflicts, provide opinions and interact with instructors and peers through dialogs [6, 7]. Existing research focuses on visualising discussion participation and social interactions [8, 9], however, analysis and the visualisation of discussion content (i.e. written discourse) is lacking. Furthermore, there is no support from existing MOOC models to effectively organise and visualise these data. In a preliminary work, Chen [10] and Speck et al. [11] focus on identifying and visualising topic models from online discussion platforms.

Due to the overwhelming abundance of information generated within MOOCs, it is challenging for the learners and the course staff to effectively locate and navigate information. Therefore, topic analysis from MOOC discussions is important in identifying main themes from students' discussions, supporting forum facilitators to become aware of the key themes and the amount of discussions in each theme. We have previously developed a framework for discourse analysis in the MOOC context that identifies latent discussion topics [12]. Our work connects lecture-related discussion topics with the corresponding weekly lectures, allowing course staff to visualise the discussions as clusters of lectures. We have experimented with our framework using three MOOCs and obtained promising results [12].

This paper focuses on developing an open source dashboard to visualise topics extracted from MOOC discussion contents. Our topic visualisation dashboard expects to answer two main questions important to the educators: *What are the emergent topics?,* and *What topics need more attention?*. Further, we also explore the topic distribution using additional variables such as views, votes, replies, and the degree of instructor intervention and answer the questions including '*what is the relationship between topics and views?*', '*what is the relationship between topics and votes*', and '*what is the relationship between topics and instructor replies*'. These questions have emerged from the authors' involvement in several MOOC courses and environments to explore key course management issues and pedagogical decisions. To answer these questions, we conducted

a statistical analysis using 3 popular MOOCs – *Machine Learning*, *Statistics* and *Psychology* and compared the results using the proposed visualisation dashboard.

2. BACKGROUND

Visual analytics within the educational context often facilitate educators in understanding large amount of learners' data to make inferences. Learners' data can be categorised as *system-generated* and *learner-generated*. System-generated data (also known as clickstream data) are generally analysed and visualised to predict the performance (e.g. CourseSignals [4]). Social Networks Adapting Pedagogical Practice (SNAPP) [8] visualises the evolution of social interactions among participants of online discussion forums.

Within the MOOC context, Coffrin et al. [2] visualises patterns of engagement and performance based on student types (e.g. auditor, active, qualified). Xu et al. [13] utilises visual analytics to explore the correlation between student behavior and student success. In a preliminary work, they analysed five MOOCs using a commercial visualisation software called *Tableau* and reported that there are multiple ways to be successful in a course (e.g. submitting quizzes, lecture views). While there is considerable, as highlighted above, contributing to the development of visual analytics capacity to better understand system-generated educational data, visualisation systems to understand learner-generated data (e.g. online discourse) is lacking.

ForumDash, a preliminary work by Speck et al. [11], focuses on visualising which students are contributing, struggling, or distracted in order to facilitate instructors in targeting their efforts effectively. Using three visualisation tools, ForumDash attempts to provide insights for teachers on which students contribute to most discussions (i.e. Thought-leaders), identify topic clusters to determine the popular topics, and through a 'contribution score visualisation', students' are capable of monitoring how much they are contributing to discussion forums compared to their peers. KISSME (The Knowledge, Interaction and Semantic Student Model Explorer) is a visualisation framework to analyse online discourse with the aim of understanding the nature of interactions among learners including contributions and relationships using LSA and social network analysis [14]. Chen [10] conducts a preliminary study on visualising topic models from online discussion platforms. Another existing tool of interest that takes elements of topic identification and social network analysis is 'Cohere' [15]. The authors use argument-mapping techniques to analyse the discussion posts based on some dimensions such as whether the post is an idea, question, or opinion, in measuring the learner's performance and attention. Topic Facet Model (TFM) incorporates forum posts (mainly questions) about Java from StackOverflow for topic analysis and visualisation [16].

Thus, our motivation for developing this research occurs due to a lack of an established research to produce 'labeled' topic models to analyse overwhelming abundance of MOOC discussion contents and visualisations.

3. TOPIC VISUALISATION DASHBOARD

The overview of topic analysis and visualisation is shown in the Figure 1. The process of topic analysis is briefly discussed in Section 3.1 and the full description can be found in the authors' previous works [12] (full analysis of this work is under review).

3.1 Topic Analysis

Our previous work focuses on identifying topic clusters from *lecture-related* MOOC discussion contents. For this, we have used a state of the art topic modeling technique called Latent Dirichlet Allocation (LDA) [17]. LDA is an unsupervised learning approach focusing on discovering hidden thematic structures in large text corpora. One of the issues associated with existing topic models is its inability to label the topics, limiting their usage in end-user applications such as visualisations. It is challenging to label discussion topics due to a lack of a reference source. As a solution, we proposed an automated topic labeling approach by generating candidate topic labels from course lectures. A Naïve Bayes classifier was trained to classify discussion topic into a week or set of weeks, and document summarisation techniques were applied to obtain the most suitable labels for each topic cluster. Our approach facilitates classifying and labeling the discussion threads using course lectures.

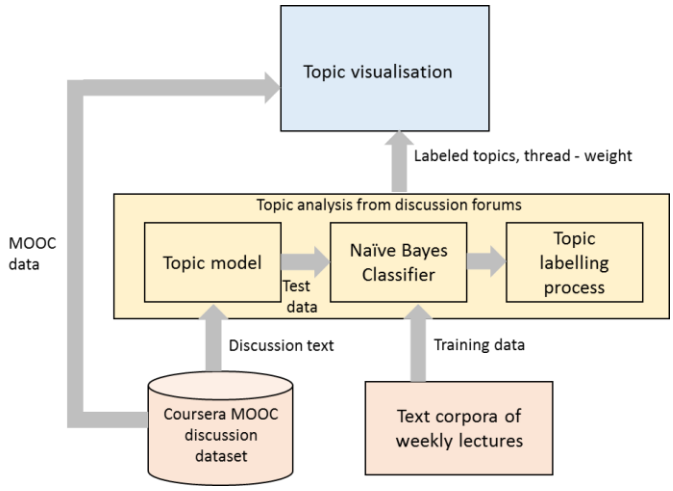

Figure 1: Overview of topic analysis and visualisation

We conducted experiments to evaluate our topic analysis approach using Machine Learning (ML), Statistics (STAT) and Psychology (PSY) MOOCs offered during 2013. In each course, we analysed approximately 5448, 2530 and 9384 number of posts and obtained 40, 25 and 40 strong topics for human annotation, respectively. Three human experts from each MOOC were recruited to label the topics manually and their mean inter-rater agreement (Kappa) was obtained as 0.75 (SD=0.09), 0.77 (SD=0.07) and 0.69 (SD=0.07) for ML, STAT and PSY respectively. We calculated the effectiveness of automated topic labeling process and obtained F-measure of 0.702, 0.75 and 0.69 for ML, STAT and PSY, respectively, demonstrating that the human-machine agreement is similar or slightly lower than inter-rater agreement. Our classifiers also performed well with a macroaveraged F-measure of 0.946, 0.926 and 0.896 for ML, STAT and PSY courses respectively We also calculated Mean Average Precision (MAP) to evaluate the ranked retrieval results of machine and obtained 0.806 (ML), 0.869 (STAT) and 0.774 (PSY). The promising results obtained from three MOOCs demonstrate that the proposed approach is effective for topic analysis of discussion contents.

3.2 Topic Visualisation

The design of our open source topic visualisation dashboard is motivated by the visual analytics process defined by Keim et al. [3] as "Analyze first, Show the Important, Zoom, filter and analyze further, Details on demand". Accordingly, our design includes analysing discussion topics, showing an overview of topic visualisation, filtering using different variables, analysing further using different variables, and providing details of individual threads on demand.

After identifying emergent topics from MOOC discussion contents (see Section 3.1), the focus of the topic visualisation is to demonstrate the discussion topics in a meaningful way for the end users, in our case the course staff, to make useful pedagogical decisions.

Our main focus is to visualise emergent topics of each course and their relationship with discussion threads. A sample screen of our visualisation dashboard using Psychology course is shown in Figure 2. As shown in Figure 2, the dashboard consists of three components; graph area, configurations, and the source.

The topic analysis is visualised by a bubble 'graph' using a force-directed layout, with larger nodes as topics and smaller nodes residing inside topics as threads. Initially, topic nodes are color-coded and adjusted in size to support visual perception of the amount of threads being discussed by the given topic (i.e. topic-thread weight). Topics are labeled using corresponding course lectures (see Section 3.1), while the similar-sized threads are initially labeled using the amount of posts associated with them. Color sliders at the bottom of the graph indicate the variations of topic-thread weight.

The 'configuration' panel (top panel of right hand side) allows the users to customise the visualisation according to their desire. Data can be imported as a CSV file for visualisation. Primarily, the data file should contain topic labels, associated thread ids, topic-thread weight and the number of posts each thread contains. However, depending on the requirement of the user, they can explore additional data such as views, votes to explore more interesting patterns. Initially, we present 10 emergent topics, supporting the visual analytics approach by Keim et al. [3] which recommends showing an overview first. The end users are allowed to adjust the number of topics up to 39, allowing a large amount of topics to be visualised for the analysis. The rationale behind limiting the number of topics to 39 is to fit into the screen resolution and similarly, if the topic-thread weight is reasonably low, it is likely that weaker topics (i.e. topic-thread weight below 0.5) are not effectively being labeled using course lectures [12]. The configuration panel also supports an optional color picker. However, the system supports variation of blue color as default.

An interesting aspect of this visualisation is that the user can explore different visualisations by changing the variables such as votes, views, instructor replies, time, number of words in threads etc. The application of the filtering parameters will change the color of topic nodes and labels of thread nodes (e.g. number of views). However, the size of the topic node remains unchanged to represent the amount of discussions associated with the given topic. For instance, number of views are vary from blue (highest number of views) to white (low number of views) (see Figure 4).

The 'source' panel provides detailed information of each thread on demand without overloading the visualisation. Users are allowed to click each 'thread' to select it and the discussions associated with this thread is shown in the bottom of the right hand side panel. In these visualisations, we have removed any identifiable data such as user or thread information.

Our open source dashboard is currently supported as a web-based system as well as standalone system which we intend to extend as a plugin embedded to the MOOC platforms.

We encounter repeated topic labels when more topic clusters are being labeled as corresponding to the same lecture. However, it is possible that these repeated topics are being discussed in slightly different threads depends on the distribution of topic terms within the topic model. If more than one topic ends up having the same label, we adjust the size of that particular topic to emphasise its' more strong influence as an emergent topic. It is also likely that a thread can be shared among multiple topics.

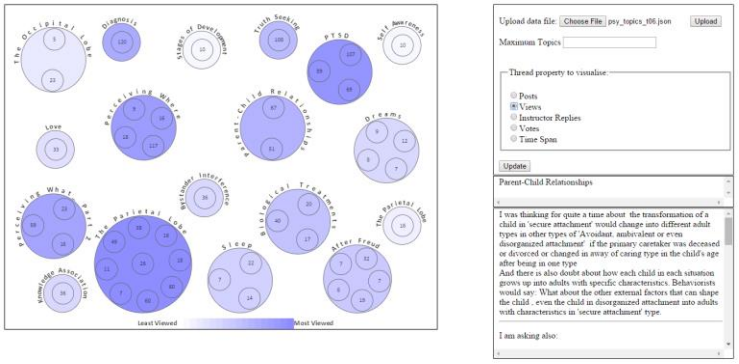

Figure 2: Topic Visualisation dashboard

It is important to determine the goals that are planned to be achieved using the visualisation in terms of improving teaching and learning within the educational context. With this in mind, we attempt to gain an understanding of online discourse at a massive scale by exploring the range of variables present in our interactive visualisation. The next section discusses our results along with interesting visualisations.

4. RESULTS AND DISCUSSION

4.1 Data

Our dataset include discussion contents (lecture-related) obtained from three MOOCs – *Machine Learning*, *Statistics: Making Sense of Data*, and *Psychology* within the Coursera platform with any user identification data removed (Table 1) [18].

Table 1. Statistics of selected MOOCs; ML-Machine learning, STAT-Statistics, PSY-Psychology

Course	Users *	Threads	Lecture-related threads	Total posts	Total words in threads	Mean (SD)
ML	6368	5449	972	5448	359,702	370 (229.6)
STAT	2313	1145	392	2530	155,329	396 (462)
PSY	11989	9300	1300	9384	719,797	553 (1014.6)

* Anonymous users are counted as 1 unit, so the number of actual discussion participants may be larger

4.2 Results

To identify emergent discussion topics of each MOOC, and as described in our earlier work, we applied Latent Dirichlet Allocation [17] and obtained 'unlabeled' topic clusters represented by a set of topic terms (usually using 10 terms). Further, we obtained list of threads associated with a given topic and their topic-thread weight (i.e. the proportion of the thread that contains the topic). From this, topics whose topic-thread weight less than 0.5 were eliminated due to the production of weak topics which mostly contain domain independent terms [12]. We apply our topic labeling mechanism to the filtered data in order to cluster the discussion topics using corresponding course lectures (see Section 3.1). Figure 3 demonstrates a sample screenshot (graph area only) obtained from our dashboard to answer the first research question in identifying emergent discussion topics.

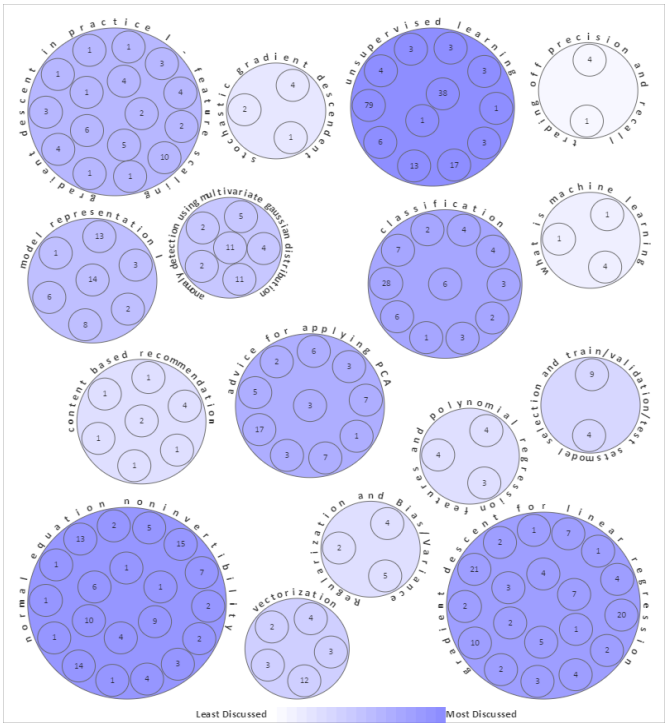

Figure 3: Sample topic-thread visualisation of Machine Learning course

As shown in Figure 3, we sized topic nodes (i.e. radius/diameter) in proportion to the number of threads to which the topic is associated, and the color in proportion to amount of posts. Even though the two topics 'content based recommendation' and 'model representation 1' are similar in size (i.e. 7 threads are associated with them), they vary in color. This occurs when the amount of posts is higher (i.e. 47) in the 'model representation 1' topic, emphasising that this topic is more thoroughly discussed by a relatively larger number of posts. The visualisation of topic-thread relationship facilitates in identifying the emergent topics as well as the topics that need teacher interventions. This visualisation of 'topic-wise classification' also assists experts in different 'topic' areas (e.g. community TAs or skilled participants) to jump into corresponding discussions and respond or assist the learners (see Figure 2 for source of thread texts). The visualisation of least discussed topics (depicted in 'lower resolution blue/white' color) assist in identification of the problematic topics for individual users or small set of users. Our approach in classifying the topics based on course lectures will also help teachers to ignore or deprioritise discussions that do not relate to course contents (e.g. social matters).

We explore topic-thread visualisation further by manipulating different variables relevant to our data including views, votes, and instructor replies to identify interesting patterns and correlations. This analysis answers the following questions;

1. What is the relationship between topics and votes?

The results of our statistical analysis show that the discussion topics and votes have a moderate positive correlation (r = 0.33; $p>0.01$) in Machine Learning course while no or negligible relationship (r = 0.13; $p>0.01$) in Statistics course since some participants tend to 'down vote' some discussions. However, Psychology course demonstrates a very strong positive correlation (r = 0.70; $p<0.01$). Thus, within the context of ML and STAT, the most discussed topics are not the ones most voted. For instance, Figure 4 demonstrates that the topic 'centre of the data and the effects of extreme values' is the most discussed topic, however, obtained only 1 vote, whereas 'data collection – observational studies' is one of those least discussed topics which obtained 8 votes. A higher number of votes suggests that the participants have more interest towards the topic or they are expecting much attention from the instructors, however, they may have less confidence to discuss it, perhaps due to lack of knowledge. Figure 4 visualises these findings.

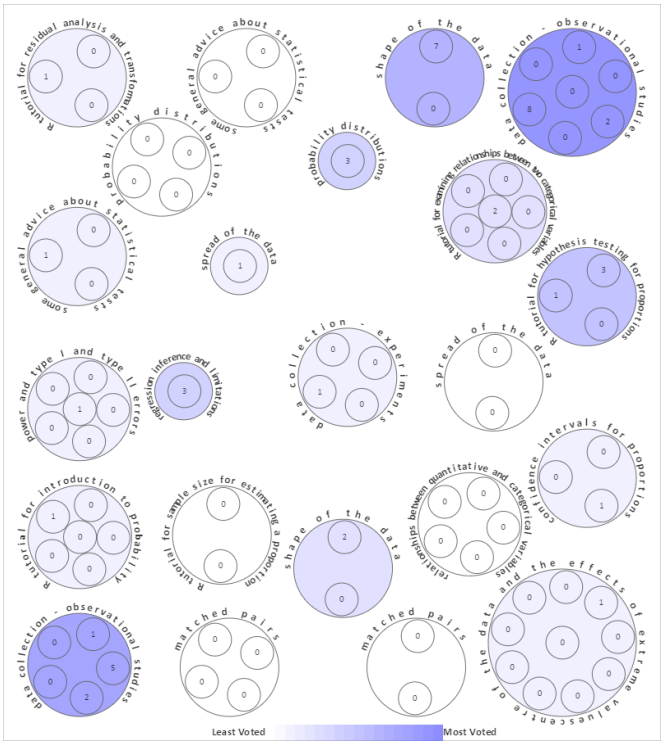

Figure 4: Relationship between topics and votes in the Statistics course

2. What is the relationship between topics and views?

We measured the correlation between the discussion topics and amount of views using Pearson correlation coefficient (r) and obtained statistically significant correlation; r = 0.7065 ($p<0.01$) for ML, r = 0.699 ($p<0.01$) for STAT and r = 0.79 ($p<0.01$) for PSY. This results suggest that the participants demonstrate more

interest towards emergent topics by viewing them more often. Similarly, less popular topics are viewed infrequently. Figure 5 depicts the visualisation correspond to this statistical analysis using the Machine Learning course.

According to the Figure 5, most discussed topics are illustrated by the size of the topic node while the most viewed topics are depicted using 'higher resolution blue' as shown in the color slider. The thread nodes are labeled using the number of views. Therefore, it is observable that the mostly discussed topics are similar to the mostly viewed topics in the Machine Learning course and vice versa. For instance, 'gradient descent for linear regression' and 'normal equation noninvertibility' are mostly discussed topics (determined by the size of the topic node) and they are also viewed more than thousand times. This kind of visualisation in classifying discussions according to topics will prioritise which posts to view and interact with based on specific requirements, resulting in a significant saving of time for both learners and teachers, particularly when reviewing massive amounts of data.

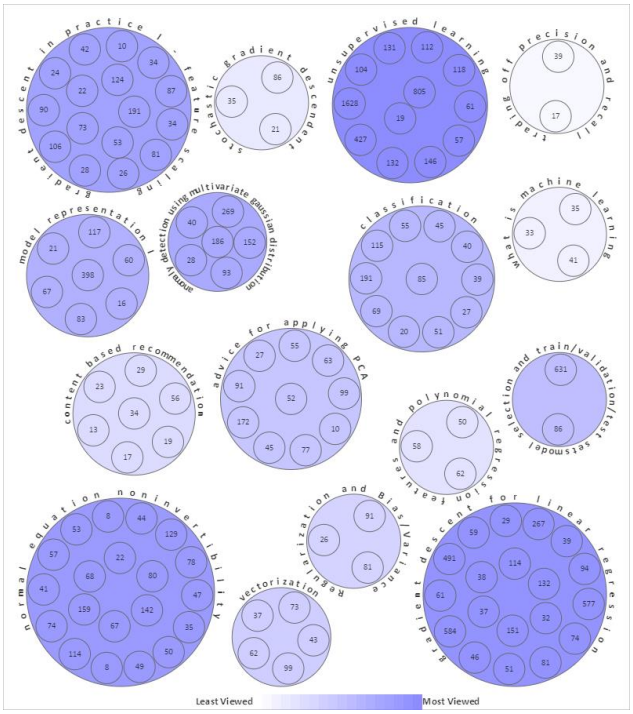

Figure 5: Relationship between topics and views in the Machine Learning course

3. What is the relationship between topics and instructor replies?

Instructor replies and discussion topics are moderately positively correlated in ML (r = 0.32; $p>0.01$). However, in STAT and PSY, these two variables demonstrate statistically significant results (r = 0.72; $p<0.01$ for STAT and r = 0.77; $p<0.01$ for PSY). This suggests that the instructors' intervention is more towards emergent topics which may isolate participants who have posted in other topics (i.e. declining topics). A study conducted by Dawson found that instructors primarily interact with high performing students despite isolated and low performing students being neglected irrespective of what they posts [8]. The ML course had relatively low instructor involvement for any topics while STAT and PSY courses had a good turnaround and strong positive correlation between these two variables. The visualisation in the Figure 6 demonstrates which topics require more inputs from instructors.

This analysis supports the open question of whether the emergent topics or declining topics require more instructor intervention. However, topic-wise classification will provide benefits to the instructors in identifying and prioritise the intervention. Simultaneously, a mechanism to 'pin' the emergent discussions will aid to avoid repeated discussions on the same topic.

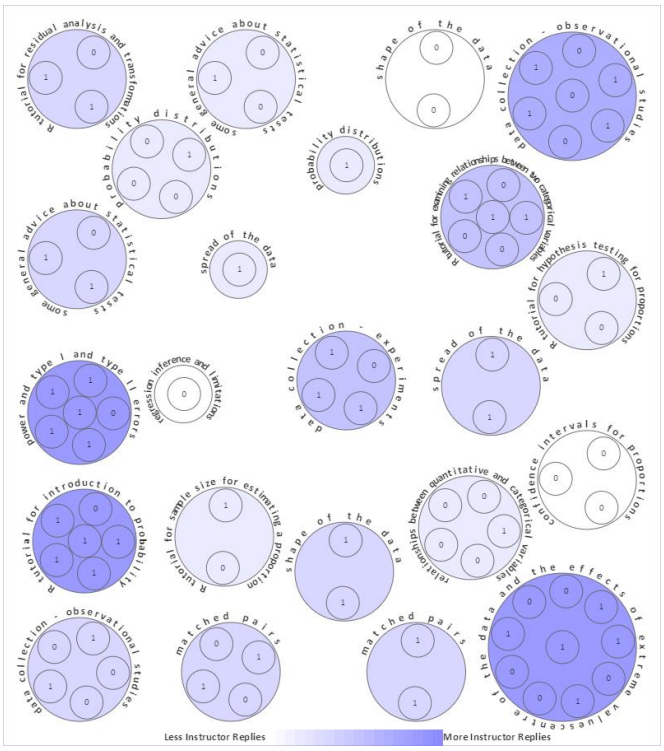

Figure 6: Relationship between topics and instructor replies in the Statistics course.

Our visualisation is currently extending to demonstrate the evolution of topics over time. The time-series analysis focuses on identifying corresponding week or set of weeks a given topic is being discussed. Some topics are discussed outside the course span (e.g. 'diagnosis' of Psychology course is discussed in week 9 where the course spans over 8 weeks). Timeline visualisation is helpful in identifying the topics that are being discussed either within or outside of the allocated weeks, enabling the identification of topics that are sustained throughout the course span.

This paper includes only a sample of visualisations and we have shared more visualisations based on the identified dataset here[3].

In summary, topic-thread visualisation assists in understanding massive volumes of discussion data by identifying emergent discussion themes, allowing the forum facilitators to make interventions more quickly rather than by reading and responding to individual threads. Similarly, topic-wise classification is supportive of comparison across discussions in understanding unexpectedly popular topics even after their expected periods in discussion.

The work presented in this paper is intended for MOOC course staff. We believe it will reduce manual forum moderation time in answering repeated questions, allowing novel discussions to occur contributing to new knowledge construction. Despite providing valuable insights into the analysis of large scale discourse, there is still considerable room for future research. These kinds of visualisation may also provide benefit to students, depending on their experience in interpreting visual information. Therefore, we consider that a topic-wise classification of discussion posts is useful as a navigational support for students, and intend to extend this work in future to support personalised navigation and recommendation of relevant posts.

This work does not yet include an in-depth analysis of individual topics or relationship between topics. It is yet to be analysed for relationship between topics and users. Our future work will include social network analysis to identify topic-inspired interactions between learner-teacher and learner-leaner (i.e. peers).

5. CONCLUSION

One of the primary challenges of MOOCs is to understand the massive volume of data to make inferences regarding student engagement or learning. To support this, our work analyses learner-generated discussion contents to identify emergent topics of discussions and labels them corresponding to the course lectures. This paper presents the visualisation of our topic-wise classification of discussion data, allowing the user to explore the analysis by manipulating different variables such as votes, views, instructor replies, and time-series analysis. A series of statistical analysis were performed to measure the correlation between discussion topics and other variables, and the finding were compared using the visualisation dashboard. This work provides benefit to the educational data mining and learning analytics research community through an open framework for topic analysis and visualisation of massive volume of discussion data generated regularly through MOOCs and other online learning platforms.

6. ACKNOWLEDGMENTS

The authors would like to acknowledge Google Inc. for supporting this project through the 'Google MOOC Focused Research Award Scheme'.

7. REFERENCES

[1] Thomas, J.J. and K.A. Cook, *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. 2005: IEEE Computer Society Press.

[2] Coffrin, C., et al., *Visualizing patterns of student engagement and performance in MOOCs*, in *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*. 2014.

[3] Keim, D., et al., *Visual Analytics: Definition, Process, and Challenges*, in *Information Visualization*. 2008, Springer Berlin Heidelberg. p. 154-175.

[4] Arnold, K.E. and M.D. Pistilli, *Course signals at Purdue: using learning analytics to increase student success*, in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*. 2012.

[5] Mercer, N., *Sociocultural discourse analysis: analysing classroom talk as a social mode of thinking*. Journal of Applied Linguistic, 2004. **1**(2): p. 137-168.

[6] Ezen-Can, A., et al., *Unsupervised modeling for understanding MOOC discussion forums: a learning analytics approach*, in *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*. 2015.

[7] Reich, J., et al., *Computer Assisted Reading and Discovery for Student Generated Text in Massive Open Online Courses.* Journal of Learning Analytics, 2015. **2**(1).

[8] Bakharia, A. and S. Dawson. *SNAPP: A Bird's-Eye View of Temporal Participant Interaction*. in *Proceeding of the 1st International Conference on Learning Analytics and Knowledge*. 2011.

[9] Oshima, J., R. Oshima, and Y. Matsuzawa, *Knowledge building discourse explorer: a social network analysis application for knowledge building discourse.* Educational technology research and development, 2012. **60**(5): p. 903-921.

[10] Chen, B., *Visualizing semantic space of online discourse: the knowledge forum case*, in *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*. 2014.

[11] Speck, J., et al., *ForumDash: analyzing online discussion forums*, in *Proceedings of the first ACM conference on Learning @ scale conference*. 2014.

[12] Atapattu, T. and K. Falkner. *A Framework for Topic Generation and Labeling from MOOC Discussions*. in *Third Annual ACM conference on Learning at Scale*. 2016.

[13] Xu, Z., et al., *Visual analytics of MOOCs at maryland*, in *Proceedings of the first ACM conference on Learning @ scale conference*. 2014.

[14] Teplovs, C., N. Fujita, and R. Vatrapu, *Generating predictive models of learner community dynamics*, in *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*. 2011.

[15] Liddo, A., et al., *Discourse-centric learning analytics*, in *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*. 2011.

[16] Hsiao, I. and P. Awasthi, *Topic facet modeling: semantic visual analytics for online discussion forums*, in *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*. 2015.

[17] Blei, D.M., A.Y. Ng, and M.I. Jordan, *Latent Dirichlet Allocation.* Journal of Machine Learning Research, 2003. **3**: p. 993-1022.

[18] Rossi, L.A. and O. Gnawali. *Language Independent Analysis and Classification of Discussion Threads in Coursera MOOC Forums*. in *Proceedings of the IEEE International Conference on Information Reuse and Integration (IRI 2014)*. 2014.