Every answer has a question: exploring communication and knowledge exchange in MOOCs through learning analytics

Irene-Angelica Chounta, Tobias Hecking, H. Ulrich Hoppe

University of Duisburg-Essen, Germany
{chounta, hecking, hoppe}@collide.info

Abstract. This paper aims to explore the use of common learning analytics methods, such as activity metrics and network analytics, in order to study and analyse the activity of users and the communication flow in discussion forums that serve Massive Open Online Courses (MOOCS). We particularly seek to identify trends and patterns that may potentially be used to support the communication and information exchange between MOOCs participants. To that end, we applied existing metrics and methods on the log files of a discussion forum that supported participants' communication for a Coursera MOOC. We present the methodology of the study as well as the results and findings with respect to knowledge exchange and information flow in the case of a massive online course.

Keywords: MOOCs, learning analytics, activity metrics, network analytics, discussion forums

1 Introduction

Over the past few years, Massive Open Online Courses (MOOCs) have gained the attention focus in the research fields of collaborative and technology-enhanced learning. MOOCs became increasingly popular after 2012 when traditional educational learning spaces shifted into online contexts [1] with examples such as Coursera (https://www.coursera.org/). The greatest benefit of MOOCs, besides openness and accessibility, is the social interaction of individual learners with large communities. Discussion forums are used to facilitate communication and information exchange between MOOCs participants within a social context. However, the massive participation makes extremely difficult the analysis and assessment of communication and information flow. In addition, it is not clear what kind of user interactions can be characterized as meaningful, thus promoting communication and, consequently, information exchange and knowledge building [2].

The main objective of this paper is to explore the use of automated and semi-automated metrics that derive from traditional learning analytic approaches in a MOOC context. In particular, we look for meaningful metrics that could potentially provide insight with respect to participation behavior in MOOCs and could be further

adfa, p. 1, 2011.

used to characterize and assess the communication, information exchange and knowledge building patterns in such learners groups. For the purpose of our study, we used the log files from a discussion board that supported a 2-months MOOC to extract metrics of user activity. In addition, we applied network analytics to trace the information flow between the participants of the course. Finally, we compared the different set of metrics in order to discover communication patterns existing in MOOCs.

2 Related Work

The term "learning analytics" is used to describe the activity of collecting and analyzing data of learners in order to understand and support the learning process. Although qualitative analysis of such data is necessary to gain a deep understanding of learning activities, data-driven approaches can provide valuable insight and suggest ways for further improvement [3]. This is particularly challenging nowadays with the explosion of big and the multi-dimensional data [4]. Plain metrics that represent activity volume, such as the sum of messages or the average number of words are commonly used to assess students' practice [5]. Additionally, network graphs and social network analysis techniques are quite popular in the field of technology-enhanced learning. Metrics from network theory, such as density and centrality, are used to assess the communication and coordination among users during learning activities [6, 7].

Online discussions, in terms of knowledge exchange, have attracted researchers over decades. One of the main goals is to define and identify expertise in such online forums. Measures of expertise can be based on the quantity of questions and answers users post to a forum or their position in the Q/A communication network [8]. However, there can be huge differences in the function and communication structure of different discussion forums. Little is known about the structure of knowledge exchange through forums in online courses although it is known that only a small fraction of registered participants on a large MOOC or discussion forums are really engaged over to complete course, completing all course activities [9]. There are often a few highly active users who have an influence on the whole community [10]. Regarding the vulnerability of the communication processes and the diffusion of information, Gillani et al. [2] have showed that it is often sufficient to take out a small amount of important users in order to interrupt the communication and information flow significantly.

3 Method of the study

3.1 Background

In this paper, we study a Coursera MOOC, named "Introduction to Cooperate Finance" that took place over a two-month period, from November 2013 to end of December 2013 [11]. The dataset provided no information with respect to the performance of users and the success of students regarding the learning objective. The Finance MOOC was supported by a discussion forum that facilitated the communication between participants. The discussion forum consisted of multiple subforums that were divided thematically (e.g. general discussion, assignments, course feedback etc.) and each subforum consisted of multiple user-created threads. A user could start a new thread by simply creating a new post. Moreover, a user had the right to comment on an existing post. The MOOC participants could post or comment in the discussion forum either using a personalized user account or anonymously. However, the anonymous users (1826 logfile entries from "anonymous" users with no further identification, such as IP address etc.) were removed from the analysis due to the fact that we aim to use personalized metrics.

3.2 Metrics and application

For the purposes of our study, we have used activity metrics that derive from the log files of the discussion forum and represent volume and ratio of activity per user. These metrics are: the number of threads that a user has submitted a post (*#threads*); the number of subforums that a user has submitted a post (*#forums*); the number of posts of a user (*#posts*); the average number of words per message per user (*wordratio*); the average sum of votes per message per user. Each user can add a (+/-) to a post (*voteratio*).

In addition, the forum posts were used to extract networks and apply network analytics metrics. For the network extraction, we classified the posts in categories according to their content. For the classification, we used a tag set classifying posts into four categories, namely "questions", "answers", "social" and "other". Other tag sets are more fine-grained differentiating between different types of questions and answers [12]. Since we aim to distinguish between active information givers and active information seekers as in [13], the reduced tag set is sufficient. As social posts, we identify posts that have a social dimensions (people looking for study groups or participants from the same country etc.) while as "others" we identify the posts that do not fall in any of the three, previous categories. Taking into account the posts classified as "questions" and "answers", we build a network of posts that link to each other. In order to build this network, we applied the following rules:

- Questions within threads are usually followed by answers. Therefore, we decompose threads as linked activity between sequences of questions and answers.
- Each post that is classified as an answer is linked directly to its parent post, provided the parent post is a question.
- An answer that is posted as a comment to a previous answer is considered to provide further information and therefore is linked to the outgoing neighbors of the parent post.

Additionally, we used similarity measures to identify lexical overlap among sequential posts. In discussion, disentanglement of chats or unstructured forums structural rules for linking contributions are often accompanied by measures of similarity between sequential posts [14]. We used lexical overlap to further refine the links between posts. The post network was further projected into a user network [15] that was used to extract the following metrics, used in this study:

- Z-score: The z-score [18] measures the deviation of the communication pattern of a forum user compared to an imaginary random poster. The z-score is defined as:
$$\mathrm{z-score(user)} = \frac{\mathrm{a-q}}{\sqrt{(\mathrm{a+q})}}$$
A user who posts more answers than questions has a positive z-score while the opposite is the case for frequent help-seeker. Users who seek for help and give help equally often should have a z-score close to 0.
- Authority score: In the sense of a directed network between forum users based on question-answer relations, a user with a high authority score would be a help seeker who receives help from many other help seekers.
- Hub score: Hubs in forums according to the HITS definition [8, 16]) are those users who give help to many other strong help-givers. A user with high hub score can be considered as extraordinary important for the knowledge exchange in the network since other help-givers rely on the user's advice.
- Inreach and outreach: The inreach and the outreach combine posting quantity and network centrality measures. Given a single node *i* in a weighted and directed network, the diversity of its in and outgoing relations can be characterised by a measure of entropy. Equation (1) calculates the diversity of outgoing relations for a node i, where $w(e_{i,j})$ is the weight/multiplicity of an edge from i to j and $od(i)$ is the out-degree i (taking into account edge weights), which is equal to the number of its help giving posts.

$$H_{out}(i) = -\frac{1}{od(i)} \sum_{j \in outneigh(i)} w(e_{i,j}) * log\left(\frac{w(e_{i,j})}{od(i)}\right) \quad (1)$$

The value for the neighbourhood entropy of node *i* reaches its maximum if all posts of *i* address different users and its minimum 0 on the other extreme. In order combine diversity and help giving activity the number of help giving posts of node i (=*od(i)*) can be multiplied with $(H_{out}(i) + 1)$ resulting in equation (2) for the outreach.

$$outreach(i) = od(i) - \sum_{j \in outneigh(i)} w(e_{i,j}) * log\left(\frac{w(e_{i,j})}{od(i)}\right) \quad (2)$$

As a result the outreach of i is at minimum the number of its help giving posts, if all posts address the same user. The formula for the inreach (help-seeking behaviour) replaces the out-degree in the formula in-degree.

4 Analysis and Results

4.1 Analysis

For the purpose of the study, we applied the aforementioned activity and network metrics on the Coursera MOOC dataset. From the dataset we removed the outliers, i.e. users who contributed no posts or that their activity could not be linked to the activity of other users or the community. Eventually, the dataset that we studied consisted of 857 participants who created 5028 posts, while the main part of the activity was spread over 31 threads and 11 forums. The 82.87% of the participants contributed less than 10 posts overall throughout the duration of the MOOC, while only a 1.28% of the participants contributed 50 posts or more. The 78.79% of the users posted in less than five different threads while only the 1.05% of the users posted in more than half of the existing threads. Finally, the 46.39% of the participants posted in one or two subforums but only 0.23% of the population participated in all subforums. The distribution of users over the number of posts, threads and subforums is presented in **Fig. 1**.

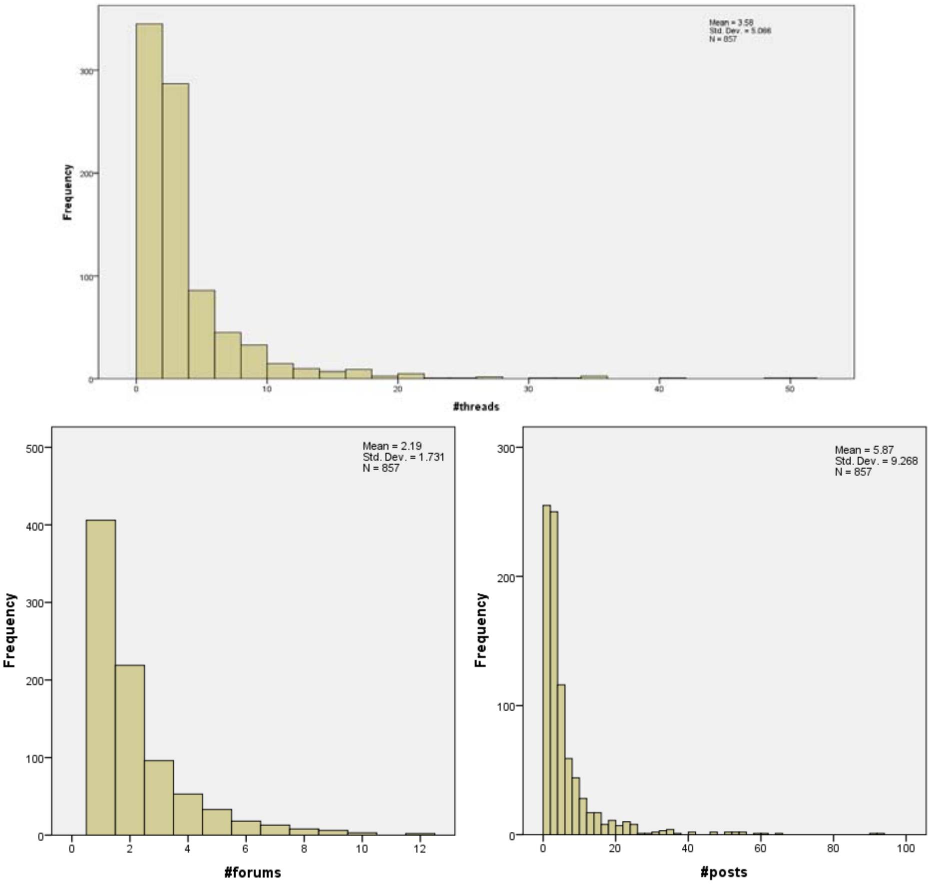

Fig. 1. Distribution of users over number of posts, threads and subforums

4.2 Results

The descriptive analysis of the dataset indicated that the distribution of users over the discussion forum was concentrated over specific threads and that the participants did not scattered over different thematic areas (subforums). In addition, it was shown that most of the participants had a small contribution, in terms of posting activity, and mostly focused on asking and answering questions. In order to explore this further, we extracted a network of users based on their posting activity and in particular, the help-giving relations between the users. On the aforementioned network we applied network analytic metrics and further studied the results in relation to the logfile activity metrics, as described earlier. To that end we used the Spearman's Rank correlation coefficient ρ since the data are not normally distributed. Regarding the activity metrics as calculated per participant, the number of threads, forums and posts that participants are active all correlate highly ($\rho > 0.8$, $p < 0.01$). Users who have a high volume of activity also spread a lot among threads and forums. The volume of activity (number of posts, threads and forums) also correlates with the number of votes received on average per user. This finding possibly indicates that the most active forum users, usually considered as "gurus" or "help-givers", and other participants use to show appreciation or preference by voting for their posts. The average number of words that participants use per post does not correlate significantly with any other metric, other than the vote ratio (on a low level, $\rho = 0.163$, $p < 0.01$). On the one hand, high posting activity does not necessarily lead to long posts. On the other hand, it is expected that well-elaborated comments, thus longer, will get more votes from other users.

In addition, we studied the distribution of the network metrics of the participants described in Section 3.2. The majority of the participants have low scores for the network metrics that were used in this study. The distribution of the inreach and outreach is similar, showing that the users of the forum are asking questions and providing answers equally. Moreover, most of the users portray low authority and hub scores. This means that they do not contribute significantly in the knowledge exchange and they do not have a key role in the communication and information flow. This finding comes in agreement with the results from the activity metrics analysis and confirms that the majority of the participants might be perceived as circumstantial users of the discussion forum that support the course, rather than exchanging information and communicating.

Some of the high correlations among the network measures and between the network measures and the z-score are not surprising. Inreach and authority score as well as outreach and hub score both increase with the number of ingoing and respectively outgoing connections. Therefore, the correlations between this measures are extraordinary high. However, between inreach and outreach there is no and between authority score and hub score there is only little correlation ($\rho = 0.222$, $p < 0.01$). Consequently, the users in the forum can be seen either as help givers or help seekers but not both in most cases. The high negative correlation ($\rho = 0.78$, $p < 0.01$) between inreach and z-score results from the fact that help seekers with more question post than answer posts have a negative z-score but a higher inreach. More interesting are the statistically significant correlations between the network measures and the other activity metrics.

Outreach correlates higher with the number of threads and forums compared to inreach. This could mean that users who are help-givers spread information in a more diverse way while help-seekers tend to ask their questions in dedicated threads and sub forums. There is also a slight correlation between outreach and votes ratio (ρ= 0.166, p<0.01) as well as between hub score and vote ratio (ρ= 0.127, p<0.01). Consequently, help givers are more likely to receive votes for their posts than help seekers.

5 Conclusion and future work

This paper presents the application of common learning analytics in MOOC discussion forums. The main objective of the study was to track and study the interactions among participants of a 2-months MOOC with respect to the communication and the knowledge exchange between them. A discussion board was used in order to facilitate the communication of users. We used the user data from the discussion board in order to build a network based on user interactions (questions – answers, posts - comments) and then we applied user activity metrics and network analytics. From the descriptive analysis of the activity metrics, it was shown that the majority of posts in a MOOC discussion forum can be classified as questions & answers (63.76%) while the "social" posts are considerably less (14.70%). Furthermore, it was evident that the majority of participants were not particularly active, with about 82.7% of the users contributing less than 10 posts in a two months period and keeping focus on certain threads. However, the participants with high posting activity, also spread among threads and subforums. These participants are identified as "help-givers" since they provide more answers than questions and get voted more from the rest of the community. The structure and the placement of the node, i.e. to whom the particular participant provides answers, is crucial and reveals the need to motivate users to contribute more.

These findings pinpoints the vulnerability of communication between participants in MOOCs as well as the inadequacy of existing tools to support knowledge exchange and promote participation on a massive scale. However, the advantage of MOOCs is partly this massiveness and the opportunity to learn and socialize with and through a large community. In order to benefit from what is considered to be one of MOOCs greatest advantages, i.e. learning within a large group and through social interaction – there is the need to motivate participants not only to contribute in discussions and information sharing but also to guide these contributions. Furthermore, the need of methods and tools to provide insights on the meaningful activities that participants should be engaged in as well as to assess the effectiveness of communication and knowledge building, is also evident. In future work, we aim to explore and study what kind of user interactions are critical for the effective communication among MOOCs participants and in what ways we can promote and assess knowledge building in MOOC communities.

6 References

1. Kay, J., Reimann, P., Diebold, E., Kummerfeld, B.: MOOCs: So Many Learners, So Much Potential... IEEE Intell. Syst. 70–77 (2013).
2. Gillani, N., Yasseri, T., Eynon, R., Hjorth, I.: Structural limitations of learning in a crowd: communication vulnerability and information diffusion in MOOCs. Sci. Rep. 4, (2014).
3. Siemens, G., Long, P.: Penetrating the Fog: Analytics in Learning and Education. Educ. Rev. 46, 30 (2011).
4. Mayer, M.: Innovation at Google: the physics of data. In: PARC Forum (2009).
5. Voyiatzaki, E., Avouris, N.: Support for the teacher in technology-enhanced collaborative classroom. Educ. Inf. Technol. 19, 129–154 (2014).
6. Hoppe, H.U., Engler, J., Weinbrenner, S.: The impact of structural characteristics of concept maps on automatic quality measurement. In: International Conference of the Learning Sciences (ICLS 2012), Sydney, Australia (2012).
7. Chounta, I.-A., Hecking, T., Hoppe, H.U., Avouris, N.: Two Make a Network: Using Graphs to Assess the Quality of Collaboration of Dyads. In: Collaboration and Technology. pp. 53–66. Springer (2014).
8. Zhang, J., Ackerman, M.S., Adamic, L.: Expertise networks in online communities: structure and algorithms. In: Proceedings of the 16th international conference on World Wide Web. pp. 221–230. ACM (2007).
9. Clow, D.: MOOCs and the funnel of participation. In: Proceedings of the Third International Conference on Learning Analytics and Knowledge. pp. 185–189. ACM (2013).
10. Wong, J.-S., Pursel, B., Divinsky, A., Jansen, B.J.: An Analysis of MOOC Discussion Forum Interactions from the Most Active Users. In: Social Computing, Behavioral-Cultural Modeling, and Prediction. pp. 452–457. Springer (2015).
11. Rossi, L.A., Gnawali, O.: Language Independent Analysis and Classification of Discussion Threads in Coursera MOOC Forums.
12. Kim, S.N., Wang, L., Baldwin, T.: Tagging and linking web forum posts. In: Proceedings of the Fourteenth Conference on Computational Natural Language Learning. pp. 192–202. Association for Computational Linguistics (2010).
13. Stump, G.S., DeBoer, J., Whittinghill, J., Breslow, L.: Development of a framework to classify mooc discussion forum posts: Methodology and challenges. In: NIPS Workshop on Data Driven Education (2013).
14. Hoppe, H.U., Göhnert, T., Steinert, L., Charles, C.: A Web-based Tool for Communication Flow Analysis of Online Chats. Networks. 11, 39–63 (2014).
15. Harrer, A., Hever, R., Ziebarth, S.: Empowering researchers to detect interaction patterns in e-collaboration. Front. Artif. Intell. Appl. 158, 503 (2007).
16. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM JACM. 46, 604–632 (1999).