The Role of Data Literacy within a MOOC Analysis

Annika Wolff, Lorraine Hudson, Gerd Kortuem
Department of Computing and Communications
The Open University
Milton Keynes, UK, MK7 6AA
{annika.wolff;l.e.hudson;gerd.kortuem}@open.ac.uk

Abstract

This paper discusses the role of data literacy in the planning of analysis of data from a six week Smart Cities MOOC delivered on the FutureLearn platform. The aim of the analysis was to discover whether the MOOC had met the aims of engaging participants with topics related to smart cities and to evaluate social interactions and understanding of the key concepts through analysis of MOOC comments. The paper identifies where data literacy impacts on decisions made, such as the need to include both domain and data expertise in the analysis, whether this is provided by a single person or by a team. It also identifies a need for better tools for rapid protoyping of methods for analysing large data sets particularly of non-standard data, such as natural language data. This would be of benefit in cases where the analysis will be used just a few times for a specific purpose, such as analysing the MOOC data across several presentations.

Introduction

This paper presents a case study of a MOOC that was developed to teach about the topic of Smart Cities and which was delivered through the FutureLearn platform. Like many MOOCS it was assembled by a small team, but with the hope that the impact of the materials could be large. This paper will explore the types of learning analytics that can be used on MOOC data and tries to understand the role of data literacy within this analysis.

There is no single clear definition of data literacy. However, the general consensus across a number of existing definitions [CAL13][MAN13][SCH5][VAH6] is that data literacy is "*the application of a data inquiry process to formulating and answering real-world questions from small or large datasets*."

The commonly cited definition of learning analytics describes it as "*the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs.*" Common applications include predicting and responding to student drop-out, or using analytics to improve future module materials for better retention [ARN12][WOL13].

Approaches to Data Literate Learning Analytics

Putting together the above definitions, a data literate approach to learning analytics is one in which the analysis of learning data is undertaken to answer the specific questions of learners and educators.

When an educator analyses data from their own classroom, they are in a position to know the sorts of insights that may be gained from this data. They have a close connection to both the data, the context in which it was gathered and from this they can understand the sorts of questions it might be possible to answer from the data. For example, they might ask which students have a low grade average and are at risk of failing the course, or they might look at patterns in attendance and correlate this to performance to

highlight to students the importance of turning up for lectures. The datasets in this scenario are typically quite small and standard data analysis techniques can be used to gain these types of insights. Therefore the educator can draw upon their own data skills for the analysis.

At the other end of the scale, where learning content is delivered online to large numbers of students the datasets can get very large. In this case, we suggest that it is much harder for an individual educator to ask and answer questions from the data. Barriers include complex infrastructures for the collection, storage and retrieval of data as well as the necessity to use more complex data analysis tools and techniques, which may be unfamiliar to an educator who is not also a data expert. For this reason, these types of learning analytics are often a team effort, drawing on expertise of data experts, as well as domain experts such as educators, course creators, or database administrators. This mix of competences is important to ensure that data literate outputs are produced, such as understanding what questions to ask of the data, understanding the meaning of all of the data attributes and the storage structure and for interpreting the outputs of the analysis. Due to the resource needed for this type of analysis, it typically focuses on producing methods that can, with appropriate adaptation, be scaled across a number of different contexts. For example, adapting a method for predicting student drop-out on one course to another similar course.

This paper explores a third scenario, in which data is collected from a presentation of a Smart Cities MOOC through the FutureLearn platform. Analysis of this data is of interest to the creator of the MOOC for the purpose of assessing the success of the MOOC in terms of its intended outcomes and to make possible changes in how the course is facilitated in the future. The requirement is to develop a set of bespoke learning analytics methods suited for a very specific context, but which must be applied across a large dataset. Whilst some of this intelligence can be derived from fairly standard data analysis, the more domain specific insights - which are potentially the most interesting - require more complex analysis of non standard data, such as natural language data. Whilst MOOCS are generally deployed across a limited number of available MOOC platforms, they are created by a diverse number of institutions. Whereas intelligence about learning analytics may be shared across an institution for the purpose of scaling up a developed approach, there is much less motivation for sharing of methods for analysing MOOCs.

Therefore, this paper describes a case where methods were developed from scratch and by a small team, comprising the course creator (domain expert) and a practitioner of learning analytics.

We will first describe how the FutureLearn course was structured, then the aims of the Smart Cities MOOC and finally the approach that was taken to developing methods for analysing the data. Finally, we will explore the role of data literacy in deciding the analysis.

Structure of a Futurelearn MOOC

Futurelearn MOOCS have a well-defined structure. Content is organised into weeks. Each week is comprised of a number of distinct learning steps, which is related to a sub-topic of the week's theme. There is a discussion space attached to each learning step where learners can post or respond to comments or questions. Each MOOC is assigned a number of facilitators who are there to help learners by facilitating conversation by posting open-ended questions and encouraging discussion between learners.. Facilitators also monitor discussions during the MOOC in order to intervene and provide help and answer questions where needed. A learner can track their overall progress in the MOOC by marking individual steps as complete.

Smart Cities MOOC

The Smart Cities MOOC premiered on Futurelearn on 28th September 2015. To date, there has been one complete presentation of this MOOC and there are 5 more currently planned, 4 in 2016 and one in Spring 2017. This MOOC has been developed as part of the MK:Smart project (www.mksmart.org). The MOOC offers learners an introduction to smart cities, providing an insight into the role that smart technologies and data can play in addressing city challenges and how citizens can get involved in their creation. The course is designed to provide the knowledge, skills and tools for learners to co-create a smart city project. Learners are prompted to consider a number of different topics throughout the MOOC, including the role of design, systems thinking, living labs, open data, crowdsourcing, finance, business models and standards in their design and development. They also join global debates on the challenges of privacy, ethics and security and weigh up the value of different approaches to smart leadership, partnerships, strategy and metrics. The Smart Cities MOOC was conducted over 6 weeks.

Analysing the Data

In the first phase of analysis, the domain expert obtained a range of data associated with the first presentation of the MOOC and applied through the appropriate channels for permissions to use the data. There were two types of data that were obtained. These are described in more detail in the following sections.

Summary Data

The first type of data obtained was Futurelearn summary data that is provided as standard to all course creators and which is made available via a dashboard accessible on the Futurelearn platform. This data is relatively straightforward to yield insight from as it is already presented in summary form. It includes:

- Course run measures - number of joiners, leavers, learners, active learners, returning learners, social learners, fully participating learners and statements sold.
- Weekly course run measures – learners visiting steps, active learners, social learners, visited steps, average visited step per users, completed steps, average completed steps per user, comments, average comments per user.
- Totals – steps visited, steps completed and comments posted.

This type of data affords the identification of overall trends, such as the attrition rate, and measures of success such as how many students signed up to the course, how engaged the students were overall by the number of learning steps they marked as complete, how much they contributed to discussions etc. These findings can easily be compared against similar data from other MOOCS, where available. This facilitates answering questions such as 'How successful was my course compared to another similar course?' or 'How does attrition rate on my MOOC compare to attrition rates for MOOCS in general?' These sorts of comparisons are useful to the course creator - in isolation MOOC attrition can appear quite high but in reality high attrition is common due to differing motivations of MOOC learners and the facility to easily stop learning in one presentation and pick it up again in the next. Therefore, comparing against other MOOCs gives a much more accurate measure of success for an individual course.

Raw Data

The Smart City MOOC team have also been granted access to additional data, which as with all FutureLearn data is provided subject to standard ethical procedures around its use. This second type of data is more fine-grained learner data and includes the comments that were made on individual learning step, students enrolled on the course, question responses and step activities. This data was collected at a cut off point of 2 weeks after the Smart Cities MOOC officially ended. At this point there was a total 7414 comments across the 6 weeks of the course. For obvious reasons, there are certain restrictions placed on how this data can be analysed and how the analysis can be reported. One particular issue is that learner-created content is published on the FutureLearn platform under a Creative Commons Licence (Attribution-Non Commercial-NoDerivs; BY-NC-ND), which means that any learner comments quoted in research publications must be attributed to the author'. However, this may be counter to the actual wishes of the participant who has commented on the FutureLearn platform but does not necessarily want to be identified in association with a comment within reports on the MOOC analysis. Therefore, analysing across comments to identify general trends is generally more appropriate.

Due to these issues, and to the fact that analysis is in very early stages, what follows is an exploration of the process of choosing appropriate analyses and eliciting questions from the domain expert, rather than an actual analysis of the data obtained. It is hoped that by the time of the workshop, it will be possible to illustrate using examples from the final data analysis.

The first point to note is that this second type of data poses challenges for analysis, it is raw data (i.e not in summary form as with the standard FutureLearn data) and it contains a lot of natural language. Therefore, unlike the summary data, it does not immediately afford the sorts of questions that can be answered from it. Some forms of analysis must be applied across it to yield insights. FutureLearn recognises this and does provide some advice and step by step tutorial for additional analyses in Excel on this data, but these answer only a fixed set of questions defined by FutureLearn and may not reflectall the questions of the course creator as was the case here.

The initial approaches proposed, such as a manual thematic coding analysis, while potentially providing detailed insight into the data are very time consuming and therefore not easily replicated multiple

times for each presentation of the MOOC, given that the comments for each MOOC represents a very large corpus of text.

A more practical approach is to develop an automated method, such as applying NLP methods to extract entities from the comment text and to verify this against the results of the manual analysis for the first MOOC presentation, with a view to developing a suite of tools for automatic analysis of later presentations. A broad analysis of discussion patterns will also be conducted, to discover whether there were any unusual patterns in commenting on individual steps of the MOOC, either eliciting more, or less, comments than other similar steps (which could be identified as dedicated discussion steps, or by type of media presented in the learning step, such as video or text-based article) or with more 'discussion' as opposed to individual comments as identified through the proportion of responses to 'new' comments.

The role of data literacy

Data literacy has impacted on the planning of the analysis in the following ways. Firstly, in combining the expertise of the domain expert with the learning analytics practitioner it reduces a risk of mismatch between the analytic techniques applies and the questions that are being asked by the educator. In this case, the primary question that the course creator has wanted to know is whether or not the MOOC has achieved the key aims of engaging participants in key smart city topics such as smart citizens, smart infrastructure, technology and data, innovation and enterprise, leadership and strategy and measurement and learning. The analysis of comments will hopefully reveal the extent to which the participants who discuss the MOOC are picking up on important concepts related to these broad topics and what are their views about this topics. As discussed previously, whilst the domain expert is able to manually undertake this analysis and interpret the outputs, it is not practical to replicate this across all future presentations of the MOOC. Therefore, the learning analytics practitioner plays a key role in bringing new technologies for analysing text data. These tools require some programming knowledge, in addition to knowledge of natural language processing machine learning even in the case that existing tools or services are being used, for example AlchemyAPI can be used for entity recognition. However, they cannot do the analysis in isolation but must call on domain expertise of the course creator to interpret and improve outputs. These sorts of conversations require some level of data literacy also of the domain expert. An alternative for the future would be to provide more accessible tools for a wider range of non-standard data analysis techniques that a non-data expert could easily incorporate into their analysis. This would reduce the need for teams of people to work on bespoke analysis methods in specialist scenarios such as analysing an individual MOOC, something which it is not always possible to resource within the budget for creating and delivering the MOOC.

Conclusions

This paper presents some very initial planning for analysis of MOOC data and identifies how data literacy impacts on this. Since the analysis is in progress, it is not currently possible to present the outputs. However, there should be results available very soon.

References

[ARN12] K.E. Arnold & M.D. Pistilli. 2012. Course signals at Purdue: using learning analytics to increase student success. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 267-270). ACM.

[CAL13] Javier Calzada Prado and Miguel Ángel Marzal. 2013. Incorporating Data Literacy into Information Literacy Programs: Core Competencies and Contents. Libri: *International Journal of Libraries & Information Services*

[MAN13] Ellen Mandinach and Edith Gummer. 2013. A systemic view of implementing Data Literacy in Educator Preparation. *Educational Researcher*, Vol 42 No. 1 pp 30-37

[SCH5] Milo Schield. 2005. Information literacy, statistical literacy and data literacy. *IASSIST Quarterly*. 28 (2/3) pp. 6-11

[VAH6] P. Vahey, L. Yarnall, C. Patton, D. Zalles, and K. Swan. 2006. Mathematizing middle school: Results from a cross-disciplinary study of data literacy. *American Educators Research Association Annual Conference,* 5.

[WOL13] A. Wolff, Z. Zdrahal, A. Nikolov & M. Pantucek. 2013. Improving retention: predicting at-risk students by analysing clicking behaviour in a virtual learning environment, *LAK13*, Leuven, Belgium.