Forthcoming in: Vondrakova, A., Brus, J., and Vozenilek, V. (Eds.) (2015) Modern Trends in Cartography, Selected Papers of CARTOCON 2014, Lecture Notes in Geoinformation and Cartography, Springer-Verlag.

Exploring Class Discussions from a Massive Open Online Course (MOOC) on Cartography

Anthony C. Robinson

Department of Geography, The Pennsylvania State University, University Park, PA, USA.

arobinson@psu.edu

Abstract. The rise of the Massive Open Online Course (MOOC) has led to its application in Cartographic education. Students in these classes generate enormous amounts of text data in the form of discussion forum posts. Here we explore the topics and geographic references found in over 95,000 forum posts collected during the 2013 launch of Maps and the Geospatial Revolution, a MOOC taught on Coursera. Using Phrase Nets, topic modeling methods, and a named-entity extraction geocoding tool, we show how students describe their use of maps, what key topics drove conversations during the class, and the geography associated with placenames mentioned in posts. These results help shed light on how novices use and understand Cartography and show how places found in discussion forum text reflect the global reach of MOOCs.

Keywords: MOOC, Cartography, Topic Modeling, Text Analysis

1 Introduction

The recent emergence of the Massive Open Online Course (MOOC) has become a transformative force in distance education. MOOCs are designed to provide class experiences that scale elegantly to tens of thousands of students working online [1]. MOOCs are normally delivered through dedicated content management platforms that host lecture videos, written/graphical content, assessment tools, and discussion forums. These platforms also capture student interactions and contributions, which form a massive dataset worthy of exploration after the course has ended. In 2013, we designed and delivered a MOOC on the fundamentals of Cartography called Maps and the Geospatial Revolution. This course enrolled over 48,000 students from more than 150 countries for its first five-week session beginning in July, 2013.

In this paper we highlight the potential for discussion archives and other qualitative contributions from students in courses like Maps and the Geospatial Revolution to reveal interesting patterns about what diverse audiences think about cartography. Students enrolled in our MOOC generated over 95,000 forum posts in more than 13,000 threads. Weekly discussion prompts urged students to focus on geospatial privacy, mapping change, mapping hazards, mapping social media, and telling stories with maps. In addition to those prompted topics, students started discussions on near-

ly every other aspect of mapping (and the course itself) that one might envision. Since this corpus is impossible for one to readily make sense of, it serves as excellent fodder for text analysis using visual and quantitative methods to uncover key topics and to explore the places that were used to explain those topics.

Here we show how visual methods such as Phrase Nets [2] can be applied to reveal how students say they use maps. We also highlight how computational methods such as latent dirichlet allocation (LDA) [3] can be used to comb thousands of posts to reveal key themes in course discussions. Finally, we use a combined named-entity recognition (NER) and geocoding approach [4] to extract and visualize the places that students mention in their discussion posts. The results of this work provide important insights for cartography educators, as they reveal key motivations for students in a globally-diverse MOOC to use, make, and understand maps. These insights may then feed into future iterations of cartography courses of all sizes and delivery modes.

2 Background

At a broad level, our research is concerned with the general problem associated with making sense out of large text collections. Visualizing text is possible through a wide range of means, including methods such as tag clouds [5], which can provide a general overview of the frequency of terms, and more sophisticated methods that use self-organized maps to develop spatialized representations to show the topics that appear in a text collection, and to expose their relative similarity across text datasets [6]. The former approach leverages a visual technique alone, with minimal computational effort required. The latter approach requires sophisticated computational methods to identify and extract patterns, before visualization becomes possible.

Our work leverages data developed by students taking the first MOOC to focus on cartographic design (hereafter referred to as the Maps MOOC). The Maps MOOC featured five weeks of lessons on the most general cartographic competencies. Students learned how to recognize spatial thinking, understand basic spatial analyses, and apply core cartographic design principles. The activities of the more than 48,000 students who enrolled to take the class were logged in a variety of ways, and the datasets that result constitute large, complex datasets in the context of distance education. Characterizing what can be uncovered from large textual datasets is a common contemporary problem for information visualization [7] and geographic visualization researchers [8], and the Maps MOOC offers a very large text dataset in the form of over 95,000 discussion forum posts. An advantage of this dataset for cartographic inquiry is that it's reasonable to expect a large proportion of this discussion to be grounded in discussions about Geography, and therefore lend itself to geovisualization research.

To begin making sense of this large and diverse textual data, we apply the use of three complementary methods in this paper. In the following sections we apply one relatively simple visual method in the form of Phrase Nets to evaluate statements students contributed regarding how they currently use maps. We then explore how the

computational approach of topic modeling through latent dirichlet allocation can be used to mine discussion forum posts to reveal major topics that students discussed. Finally, we make use of a modified named-entity extraction method to identify the placenames that students talked about in discussion assignments and to geocode those places to explore the relevant geography for major discussion themes in the class.

3 Phrase Nets

The Maps MOOC class experience began with a prompt emailed to students to "pin themselves" on a web map to develop an overall view of the people and countries represented in the class cohort. Students in this early class activity were asked to add a pin to a world map to represent their home, and to provide basic demographic information (age range and gender). Students were also asked to provide a simple, one sentence answer to the prompt, "How do you use maps?" This prompt was intended to elicit a wide range of opinions from around the world regarding how novices view Cartography prior to completing the Maps MOOC. Of the 22,781 pins added to the map, 11,710 had complete data for age, gender, and the map usage question. We focus here only on these complete observations.

To evaluate student responses to the question "How do you use maps?" in the opening map assignment for the course, we turn to a technique developed to support visual analysis of phrases. Phrase nets were conceived by van Ham et al. [2] to aid exploration and visual analysis of phrases in text. The method allows users to select relations (either syntactic or lexical) and to view what frequently comes before and after those terms. For example, in Figure 1 (Top), we show how the word "and" is used to link words commonly found in the responses we gathered from students reflecting on how they currently use maps. Links between pairings are shown with lines of varying thickness depending on the degree to which those terms often co-occur. Some of the pairings of interest include *Maps and GIS*, *Travel and Work*, and *Place and Directions*.

Changing the relation used to form the Phrase Net can reveal other interesting patterns in this dataset. If we choose "the" as the relational linkage (Figure 1, Bottom), one can see that a key phrase in our responses was *Understanding The World*. This was complemented by less frequent pairings such as *Discover The World*, *Explore The World*, and *Navigate The World*.

Fig. 1. (Top) Terms in student responses about how they use maps that are linked by the word "and." (Bottom) Terms in those same responses that are linked by "the."

Expanding the lexical relations in Phrase Nets can further reveal how maps are viewed by students in the Maps MOOC. Figure 2 (Top) shows the resulting visualization for the "of the" relation. This structure pulls out a colloquialism in *Lay Of The Land*. The most common phrase here is *Understanding Of The World*, with evidence

that significant numbers of students also wrote about *Sense Of The Place* and *Parts Of The City*, the latter of which is one of the few scale and context-specific references that appears in any of the phrase nets we developed.

Since Phrase Nets can also use syntactical relations to arrange text, we can also look at the basic pairings of words separated by spaces alone. Figure 2 (Bottom) shows the extent to which students view technology as a key aspect of their map usage, in pairings such as *Google Maps*, *GIS Data*, and *Find Directions*, among others.

These and other examples using our "How do you use Maps?" dataset can be further explored using IBM's ManyEyes Phrase Net tool at http://ibm.co/1f8B2EO.

Fig. 2. (Top) Terms in student responses about how they use maps that are linked by the word "and." (Bottom) Terms in those same responses that are linked by "the."

4 Topic Modeling

Topic modeling is one popular computational approach for analyzing text data. Specific techniques for topic modeling include basic probabilistic methods that predict the likelihood that one word follows another [3], and somewhat more sophisticated methods such as latent dirichlet allocation (LDA) which can model topics independent of word order [9]. Many options exist today for alternative approaches which advance upon these basic examples, with new combinations and modifications appearing all the time. LDA, however, has remained a popular method for topic modeling, and a large number of tools are available today for researchers to apply which leverage the LDA approach. The Machine learning for language toolkit (MALLET) is one such example that uses LDA to mine topics from text [10]. MALLET is built using the Java programming language and provides command line controls for processing large text collections to extract key topics using LDA. Since its first iteration in 2002, MALLET has been improved in several stages, and the tools now include methods for tagging sequences and classifying documents, among others.

To make MALLET easily usable by non-experts, David Newman at the University of California-Irvine created the Topic Modeling Tool (TMT) to provide a graphical user interface to MALLET (http://code.google.com/p/topic-modeling-tool/). We used the TMT in our work to reveal key topics found in discussions in the Maps MOOC. Input data for TMT was comprised of the text from 95,958 discussion forum posts created by students in the class.

Figure 6 shows an overview of the top 20 topics identified by MALLET from our discussion post dataset. We have further categorized the top 20 topics into three key themes that appear to link individual topics; Mapping Technology, The Course Itself, and Geography. Students frequently discussed the impact of mapping and location technology, with a specific interest on changes in the ways maps are made today. One topic shows artifacts of a very common URL shortening service, which was widely used by students who shared the web maps they created with one another.

The second major category of topics concerns the course itself. Students frequently talked about class policies, their goals for taking the MOOC, and aspects of assignments that they struggled to complete or understand. One outlier here is the final listed topic in this category which shows several Spanish words and references a wiki. A large group of Spanish-speaking students took this course, and the Spanish-speaking study group thread in the study group forum was among the most active of all study groups. Students in that thread and others frequently posted links to Wikipedia articles in Spanish to help elaborate concepts from the lectures in the course.

Finally, the largest category of topics concerns Geographically-focused discussions. Students talked about the discoveries they made about populations, land cover change, hazard analysis, and social media – all of which were key themes in their lab assignments for the course. Urban geography was of particular interest, as were topics centered on change of all types. One of the first discussion prompts in the course asked students to consider spatial privacy concerns, which resulted in a few related topics shown here in this category.

Fig. 3. Major topics uncovered from discussion forum posts using MALLET.

5 Geo-Parsing Forum Posts

We continued our exploration of text generated in the Maps MOOC by applying text mining techniques to identify and geocode the place names mentioned by students in forum posts. This general process is frequently called geo-parsing in contemporary literature [11]. To analyze our course forum dataset of over 95,000 posts, we used the GeoTXT.org service. GeoTXT combines named-entity recognition methods with geocoding capabilities in order to extract and locate placenames found in text [4].

Processing the forum post dataset with GeoTXT resulted in the extraction of 38,317 places from 95,958 posts. There were 21,357 posts that included at least one place mention. Some posts referenced multiple placenames; an overview of which can be found in Figure 6.

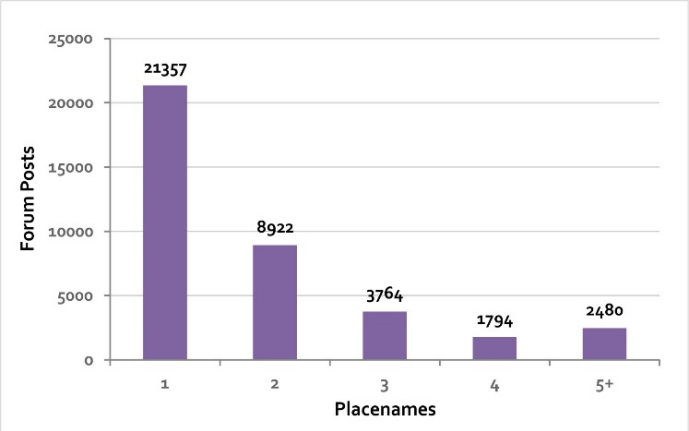

Fig. 4. Chart showing the number of placenames found in MOOC forum posts.

After extracting and geocoding the placenames mentioned in class discussion forum posts, we sought ways to map these data in order to understand the frequency and distribution of placename mentions. First, we aggregated all placenames found into hexagons (2 degrees wide at the Equator), in order to explore the overall global density of place-oriented discussion in the class. Figure 7 shows the resulting map, which looks much like a population density map of the Earth, with some notable exceptions in China and parts of Africa.

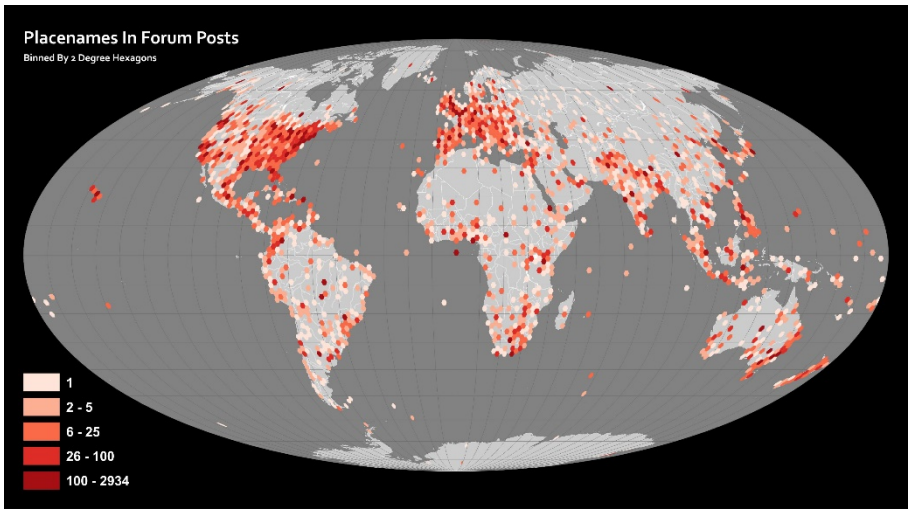

Fig. 5. Map of placenames found in MOOC forum posts, aggregated by 2 degree wide hexagons.

While ordinarily it would be inappropriate to use a non-normalized choropleth map to show these data, in this instance, the raw totals aggregated by country helps show

the general pattern of placename mentions in the class forum that is less apparent in the hexagonal binning results. Many of the hex bins are located on the centroids of countries, indicating the mention of a country name in a discussion forum post. Figure 8 shows the raw totals aggregated to countries, revealing a not too surprising trend around highly populated, English-speaking countries. The map does not track overall population trends when one considers Asia and Africa.

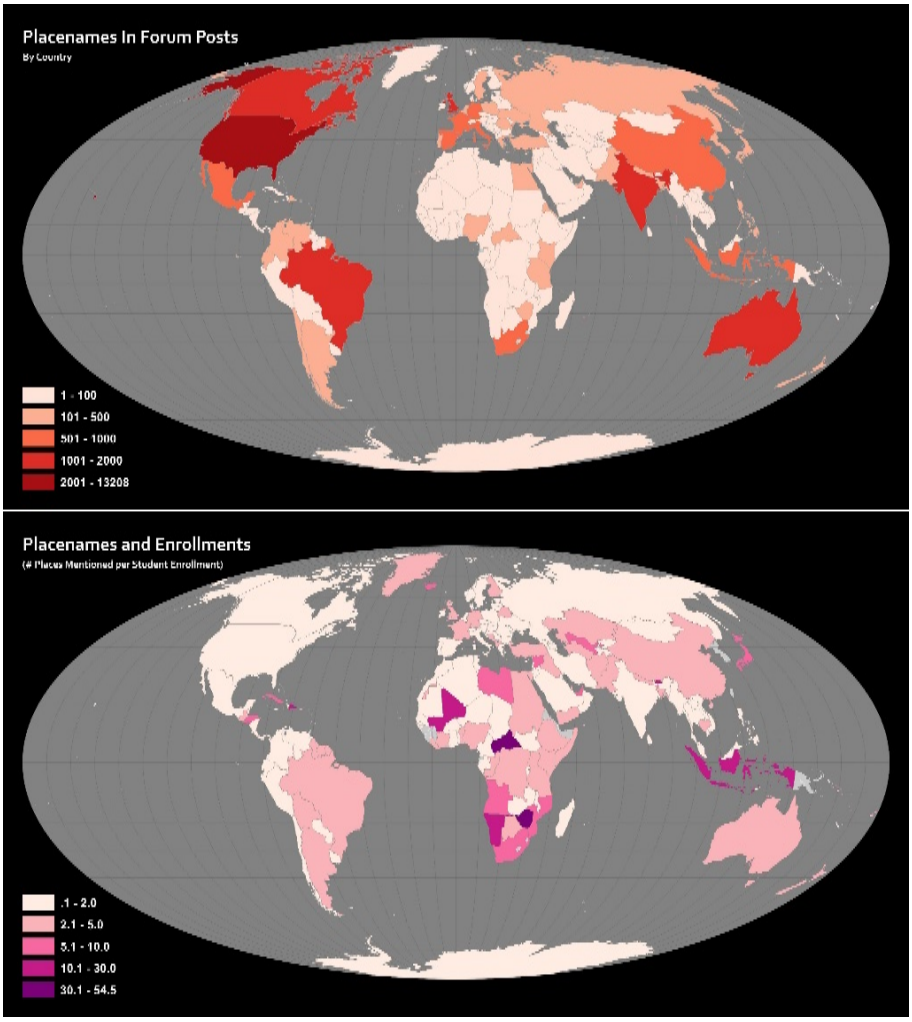

Fig. 9. Map of placenames found in MOOC forum posts; raw counts shown in the top map, and counts normalized by the number of students enrolled from each country in the bottom map.

These analyses prompted us to think of a way to normalize the data by a reasonable measure. We found overall population normalization was not particularly interesting, as it simply emphasizes very small population countries that received a handful of mentions (Greenland and Antarctica, for example). Instead, we used the "Pin Yourself" mapping activity data from the beginning of the class to count enrollments by country, and then used those enrollment data to normalize the placename mentions in our forum data. The resulting map is shown in Figure 9. What is clear from this map is that there are some places that are of greater interest in discussion than they were in terms of attracting students, and that this is particularly true in Africa and parts of Asia. Some countries, like Brazil and Australia, were talked about in discussion frequently, and also constituted sizable proportions of the overall student population (both appearing in the top ten placename and student enrollment lists).

6 Conclusions

In this paper we have contributed examples of visual, computational, and combined visual-computational analysis of forum data from a MOOC on Cartography. Our results reveal the ways in which novice students perceive the utility of maps, the key topics students discussed during the course itself, and the Geography embedded in those forum discussions.

There remains much work to do to further explore and analyze these data. Of specific interest to us is the relationship between time and the analyses we have shown here. For example, discussion topics changed over time from week to week in the class, and one wonders whether or not the placename references may vary along with those changing topics as the course progresses. We can also further explore the extent to which where a student comes from has an influence on the places they mention as examples in topic areas like spatial privacy, natural hazards mapping, and social media mapping (all of which were discussion themes prompted by the instructor in this course).

Sentiment analysis is another fruitful potential direction. Students in MOOCs are expected to help each other, given the fact that no single instructor (or team of instructors) could possibly interact with each student individually in a significant way. Some students find this frustrating, while others take the opportunity to self-organize into study groups to help one another. We hypothesize that there are likely geographic differences in students who choose to organize versus those that do not, and that student sentiment may vary accordingly.

Ultimately there remains much left to do with these data beyond the basic analysis we have shown here. A larger future goal should be to support rapid analyses of MOOC data to uncover spatio-temporal patterns of student activity and interest may serve as critical aids to instructors who are trying to adapt their class content and assessments to an ever-changing, global population of students.

Acknowledgements

We thank Jan Oliver Wallgrün for assistance in using GeoTxt.org to geo-parse the forum data analyzed in this paper.

References

1. McAuley, A., Stewart, B., Siemens, G., Cormier, D.: The MOOC model for digital practice. (2010)
2. van Ham, F., Wattenberg, M., Viegas, F.B.: Mapping text with phrase nets. IEEE Transactions on Visualization and Computer Graphics 15, 1169-1176 (2009)
3. Wallach, H.M.: Topic Modeling: Beyond Bag-of-Words. 23rd International Conference on Machine Learning, pp. 977-984, Pittsburgh, PA (2006)
4. Karimzadeh, M., Huang, W., Banerjee, S., Wallgrün, J.O., Hardisty, F., Pezanowski, S., Mitra, P., MacEachren, A.M.: Geotxt: A web API to leverage place references in text. 7th ACM SIGSPATIAL Workshop on Geographic Information Retrieval, Orlando, FL (2013)
5. Hassan-Montero, Y., Herrero-Solana, V.: Improving tag-clouds as visual information retrieval interfaces. International Conference on Multidisciplinary Information Sciences and Technologies, pp. 1-6, Merida, Spain (2006)
6. Skupin, A., Fabrikant, S.I.: Spatialization methods: A cartographic research agenda for non-geographic information visualization. Cartography and Geographic Information Science 30, 99-119 (2003)
7. Dörk, M., Gruen, D., Williamson, C., Carpendale, S.: A Visual Backchannel for Large-Scale Events. IEEE Transaction on Visualization & Computer Graphics 16, 1129-1138 (2010)
8. MacEachren, A.M., Jaiswal, A., Robinson, A.C., Pezanowski, S., Savelyev, A., Mitra, P., Zhang, X., Blanford, J.: SensePlace2: Geotwitter Analytics Support for Situation Awareness. IEEE Conference on Visual Analytics Science and Technology, Providence, RI (2011)
9. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993-1022 (2003)
10. McCallum, A.K.: MALLET: A machine learning for language toolkit. (2002)
11. Gelernter, J., Zhang, W.: Cross-lingual geo-parsing for non-structured data. 7th ACM SIGSPATIAL Workshop on Geographic Information Retrieval, Orlando, FL (2013)