Modeling and mapping thematic places

Benjamin Adams and Grant McKenzie

Abstract:

While the notion of *place* is a central concept in many humanistic studies, its inherent fuzziness means that it does not lend itself well to formal computational models. As a result places as such are not often explicitly represented in a GIS. Given the importance that place plays in human communication, however, we propose that one window into learning such representations is to statistically extract the spatial heterogeneity of the themes that people write about. Spatial heterogeneity is the notion that "geographic phenomena do not oscillate around a mean, but drift from one locally average condition to another" (Goodchild, 2009). We posit that from a corpus of natural language documents where each document is associated with a location, we can find the local regions where certain themes are most salient.

A vast number of natural language documents on the web and elsewhere are associated with one or more locations and/or times. Sometimes these documents are explicitly tagged with a location (e.g., travel blog entries) but other times the locations are implicit in the text. In the latter case named entity recognition can be used to extract the locations. The spatial and temporal ordering of these documents allows us to perform analysis on their contents to extract spatio-temporal thematic patterns. In the work we present here we specifically explore the use of a topic-modeling algorithm called latent Dirichlet allocation (LDA) to model the latent topics in geographically embedded natural language text. LDA models each document as mixture of topics and each topic as probability distribution over words. It is a generative model in that it describes how the documents are generated from these topics through a random process. In addition, it is a bag-of-words model, which means that syntactic structures in the documents are ignored. Despite these simplifying assumptions, in practice LDA provides a means to learn the latent topics in documents in an entirely unsupervised manner.

In this talk we will present some of our results from applying LDA to a set of approximately 275,000 travel blog entries. Each entry is tagged with a geographic location and date, and we examined the "hotness" and "coldness" of the latent topics over space and time. The LDA results for each entry were aggregated based on their location. We created a quarter degree grid over the world and assigned topic values to a point at the centroid of each grid square based on the average topic values of all entries found within that grid square. In order to identify places associated with specific topics we then applied kernel density estimation over these points to create a field representation of the topics. We present how this field can be visually represented on a map both as polygonal regions indicating a threshold of thematic salience and with 3D visualizations of "topic ranges".

Goodchild, M.F. "What Problem? Spatial Autocorrelation and Geographic Information Science," *Geographical Analysis* 41:4, 411—417, 2009.