Chapter 12
Inferring Thematic Places from Spatially Referenced Natural Language Descriptions

Benjamin Adams and Grant McKenzie

Abstract Places are more than just a location and spatial footprint. A sense of place is the result of subjective experience that a person has from being in a place or from interacting with information about a place. Although it is difficult to directly model a person's conceptualization of sense of place in a computational representation, there exist many natural language data online that describe people's experiences with places and which can be used to learn computational representations. In this paper we evaluate the usage of topic modeling on a set of travel blog entries to identify the themes that are most closely associated with places around the world. Using these representations we can calculate the similarity of places. In addition, by focusing on individual or sets of topics we identify new regions where topics are most salient. Finally we discuss how temporal changes in sense of place can be evaluated using these methods.

12.1 Introduction

J. Nicholas Entrikin (1991) has written that narrative accountings of places are essential resources for understanding the world because they provide "a distinct form of knowing that derives from the redescription of experience in terms of a synthesis of heterogeneous phenomena." A key aspect of these narratives is that they come from an individual point of view and therefore capture qualities of subjective experiences. Much volunteered geographic information (VGI) on the Web comes

B. Adams (✉)
Department of Computer Science, University of California, Santa Barbara, CA, USA
e-mail: badams@cs.ucsb.edu

G. McKenzie
Department of Geography, University of California, Santa Barbara, CA, USA

D. Sui et al. (eds.), *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*, DOI 10.1007/978-94-007-4587-2_12,

in the form of unstructured, natural language descriptions of places on the Earth. Examples of these kinds of descriptions include Wikipedia articles, travel blog entries, and entries from microblogs such as Twitter. VGI place descriptions form rich datasets for geographic analysis; however, the vast quantity of available information begs for automated approaches to aid analysis. In this chapter, we describe results of using topic modeling, a popular natural language-processing technique, to identify the latent topics in a large corpus of travel blog entries that describe places around the world. We examine how the topics are distributed over space and time and how individual or combinations of topics can be drawn on a map to represent places of thematic distinction.

Geography has traditions in both thematic and regional analysis. Thematic geography examines the commonalities and the differences between geographic structures through the lens of a particular theme, e.g., economics or politics. Regional geography focuses on a particular region of the Earth and takes into account the unique spatial organization of that region. 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). It is an important concept in geographic information science since it means that statistical methods that treat the world as flat will fail to accurately model many geographic scale phenomena. In a large corpus of natural language documents, where the documents are associated with one or more locations, the distribution of topics will be spatially heterogeneous. Some common thematic patterns will be found across the documents that span different geographic regions rather uniformly and other themes will be found that are highly spatially and temporally correlated with particular locations and times. Consequently, in these sorts of documents, there is grist for both thematic and regional geographic analysis.

In this chapter, we describe a method for using topic modeling on georeferenced natural language text to construct regions of thematic saliency. Topic modeling is an automated data-mining technique that has enjoyed popularity as an effective way to discover the latent topics in a large corpus of documents (Blei and Lafferty 2009). Informally, a topic is a semantically coherent grouping of terms that tend to co-occur in a document. For example, a topic might be characterized by the words: *wine, vineyard, tasting,* and *cheese*. A corpus of travel blog entries is used as an exemplar in this chapter, but the techniques presented can be adapted to other texts that are ordered in this manner. The text of each blog entry is modeled as a mixture of topics produced through a random generative process and we generalize those results over all the entries in a location. The result is that we can develop dynamic statistical formal models of places out of highly unstructured volunteered data. We show that some of the resulting topics correspond to specific geographic locations and thus are applicable for predictive analytics of the form "Where (or when) is this text about?" whereas other topics provide a means for thematic and comparative exploration.

One result of applying topic modeling to a large corpus is that it effectively reduces the dimensionality of the topic space allowing researchers to identify and compare geographic contexts based on a fixed number of themes. This reduction of

the corpus to an interpretable number of thematic dimensions creates potential for different sorts of analysis. For example, as we show later in the chapter, the topics discovered in the travelogues can be used to discern whether places are generally viewed as natural places or rather dominated by descriptions of human made features. In addition, by looking at the most prominent topics for a given place, researchers can identify the landmarks, features, and associated activities that are most salient in a place from the tourist's perspective, which can be compared and contrasted with other data about the feature distributions at a place or descriptions written by locals. They can also be used to find places that otherwise might be very dissimilar but are analogous with respect to specific thematic dimensions. Finally, when looking at a corpus of travelogues that spans over time, researchers can use this methodology to better understand how the touristic image of a place has changed over time.

The fundamental premise behind our approach for creating thematic regions is that a natural language document that describes a place is an observation of phenomena at a particular location, $\mathbf{x}$, and the mixtures of topics that compose these kinds of observations will be spatially autocorrelated. This can be viewed as a rewording of Tobler's first law that near places are more similar than far apart places (Tobler 1970). Note, that while the *mixtures* of topics will show spatial autocorrelation, the individual topics that make up those mixtures will have differing degrees of global spatial autocorrelation. In other words, some topics are more local than others.

The remainder of this chapter is organized as follows. In Sect. 12.2, we present background information on topic modeling, place, and related work on using topic modeling to find regional topics. In Sect. 12.3, we present the data collection and preprocessing process. Section 12.4 shows the results of running latent Dirichlet allocation (LDA) on the data and details the method to describe and analyze places from these results. In Sect. 12.5, we show how the topics generated can be visualized and how the regional extent of topics can be mapped, and in Sect. 12.6, we look at temporal analysis of the themes. Finally, we conclude with future research.

12.2 Background

In this section, we provide background information on place, topic modeling, and related work.

12.2.1 Place

According to Tuan, place is space infused with meaning, i.e., a way of making sense of the world (Tuan 1977). Another commonly cited definition of place by Agnew is that it is the combination of location, locale, and sense of place (Agnew 1987). By this definition, sense of place is subjective and is a product not only of the physical

structure of a place but also the phenomenological experiences that an individual has when in a place or when observing a reference to a place (e.g., reading about it) (Cresswell 2004). The conceit of the methods presented in this chapter is that what people choose to write about a place reflects their sense of that place, and by generalizing over many people's writings, we can extrapolate an aggregate view of a place. By doing this kind of analysis, we remain agnostic on the question of whether to approach the study of place from a relatively decentered and objective perspective or relatively subjective perspective (Entrikin 1991). Because topic modeling operates on the level of individual documents, analysis can be performed on the level of aggregations of descriptions (as we do below), or it can be done on the level of individual descriptions reflecting a more individualized notion of place.

Despite the importance of place in geography and related disciplines, there has not been much success in formally modeling sense of place in geographic information systems. The operationalization of place has, however, been identified as an important research agenda; notably, an issue of the journal *Spatial Cognition and Computation* was dedicated to this question (Winter et al. 2009). Multidimensional measures of sense of place have been explored in human geography, but they tend to be tested using psychological experimental studies of very specific geographic settings, e.g., lakeshore properties (Jorgensen and Stedman 2006). The goal in this chapter is to explore unsupervised ways of operationalizing place using a much larger, crowdsourced dataset generated by many different people.

12.2.2 Topic Modeling

Generative topic modeling encompasses a suite of unsupervised data-mining methods for uncovering the semantic structure of textual documents in a large corpus (Steyvers and Griffiths 2007). A generative topic model is a statistical model that explains how the words found in documents are generated as the result of a random process. For the most popular generative topic model, LDA, each word in a document is chosen from one of a set of topics that are shared among all the documents in the corpus (Blei et al. 2003). Each document is modeled as a unique mixture of topics (i.e., a multinomial distribution over topics), and each topic in turn is a multinomial distribution over words. Therefore, to generate each word, one can imagine two weighted dice being tossed. The first die has as many sides as there are topics and is weighted uniquely for each document. It is used to probabilistically sample a topic. Then, given the selected topic, we toss another dice specifically weighted for that topic and with as many sides as there are words in the corpus, drawing the word. This generative process is easily extended and can take into account other information, such as authorship, which has lead to a number of variants of LDA (Steyvers et al. 2004). Topic models are *bag-of-words* models, which means that the order and syntactic context of the words in the text do not factor in the result.

The goal of topic modeling is to infer the latent variables (i.e., the weights on the dice) most likely to have generated the observed words in a corpus of existing documents. This inference is a Bayesian-inferencing problem on a large probabilistic

graphical model. The main innovations in topic modeling over the last decade have been to develop efficient algorithms for approximating this inference, given that an exact solution to the problem is computationally intractable. Algorithms that use a Gibbs sampling Markov chain Monte Carlo (MCMC) approach have proved effective to approximate parameters (see Griffiths and Steyvers 2004; Bishop 2006).

When running an LDA inference, the inputs are the α and β hyperparameters, the number of topics, and the observed data. The α hyperparameter determines how many topics are assigned to a given document (a very small α will essentially assign one topic to every document). The β hyperparameter determines whether the words are distributed more or less evenly over the topics. A number of excellent implementations of LDA based on MCMC are freely available. For the work presented in this chapter, we used the topic modeling toolkit contained within the MAchine Learning for LanguagE Toolkit (MALLET) (McCallum 2002).

LDA is very modular, and several extensions to LDA have been developed, including ones that allow for learning the number of topics, supervised labels, and correlations between topics (cf. Blei and Lafferty 2006; Teh et al. 2006; Li and McCallum 2006; Blei and Mcauliffe 2008). Extensions add to the computational complexity of the approximate inferencing, however. In the analysis presented in this chapter, we utilized the original LDA model, though the post hoc spatial analyses of the topic modeling results presented here are fully compatible with any of the many topic modeling variants.

12.2.2.1 Similarity of Documents

The Kullback-Leibler divergence, D_{KL}, (also known as the relative entropy) of the topic distributions of two documents can be used as a similarity measure. Let P and Q be probability distributions of a random discrete variable:

$$D_{KL}(P \mid Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

Kullback-Leibler is an asymmetric measure; i.e., the distance from P to Q is different than the distance from Q to P. If a symmetric measure is desired, then the Jensen-Shannon divergence, D_{JS}, can be used instead:

$$D_{JS}(P \mid Q) = \frac{1}{2} D_{KL}(P \mid M) + \frac{1}{2} D_{KL}(Q \mid M),$$

where $M = \frac{1}{2}(P + Q)$.

12.2.3 Related Work

A location topic model for travelogues has been developed that explicitly decomposes documents into local and global topics (Hao et al. 2010). There has been some

work on extending the LDA model to include information on document labels or links, which can be used to train for topics that are predictive of a variable (e.g., a location label) (Wang et al. 2007; Blei and Mcauliffe 2008; Chang and Blei 2009). In addition, models have been developed to specifically train for spatiotemporal themes (Mei et al. 2006). However, by specifically training for topics that are predictive of location or time, we lose the ability to examine how the spatial and temporal distribution of individual topics *differ* from one another. One goal of the work presented here is to be able to characterize the degree to which different locations share or do not share themes. By focusing on post hoc analyses of topic modeling results, we create techniques that are applicable to a wide variety of source data and are not be overspecified for a specific domain. However, all the above approaches are compatible with the work presented here. In addition, although the methodology presented here focuses on text analysis, it can be augmented by incorporating other kinds of data such as images (Serdyukov et al. 2009).

12.3 Data Preprocessing

There are a number of blogging sites on the Web that allow people to post blog entries about their travel experiences. Since we were interested in exploring a diverse set of entries from around the world that were written by a variety of authors, we looked at some of the larger sites, including travelpod.com, travelblog.org, and travellerspoint.com as sources of data. Travelblog.org[1] was chosen given its relatively simple user interface, and a Web crawler was written to download the text for all the public entries through September 2010. In addition to the text, the date and location, in the form of a geographic hierarchy of the entry, were also saved. Travelblog.org entries often have images and video as well, but since we were only interested in textual analysis, we did not download those; however, such information could be used in the future for a more complex semantic analysis of the entries. In total, 309,683 blog entries were downloaded.

The entries were preprocessed for LDA using the following steps. First, some blog entries consist almost entirely of pictures or video, so entries with fewer than 100 words were removed. Second, during our exploratory analysis, we discovered that the words from blogs written in languages other than English tend to be organized by LDA into their own topics. To mitigate this problem, a language detection script was run on each entry, and those entries labeled as non-English were removed. However, due to the presence of some entries that were written in English as well as in another language, some of this effect was still found. Third, the words in the entries were filtered against a standard list of English stop words, and all punctuation and HTML markups were removed. After preprocessing, the input dataset consisted of 275,468 blog entry documents.

Travelblog.org lets users specify the location associated with an entry within a geographic hierarchy (e.g., North America, United States, California, Los Angeles).

[1] http://www.travelblog.org

Users usually select the location from a predefined taxonomy, but they can suggest a new location that will be added to the database dependent on moderator approval. There does not appear to be a standard method for determining how countries are subdivided into regions. Some countries such as the United States and France have regions based on first-level political administrative units, but other countries are subdivided into nonpolitical geographic regions or skip directly to local town/city-level regions. Users have flexibility to specify location at any depth of the geographic hierarchy, which means that some entries are labeled in a coarse-grained manner (e.g., California). In addition, it is difficult to compare regions at the same level because they reference areas that vary greatly in size. For example, Andorra and Russia are both at the same level.

We mapped the user-defined locations to geometric representations to aid the visualization and spatial analysis. One mapping was achieved by geocoding each unique location to a latitude-longitude point using the Google Maps geocoding Web service. The service handled all but about 500 locations, which were hand coded. In total, each entry was mapped to one of 10,496 locations. The coarse granularity of some locations, while problematic, is unavoidable as entries about a person's travel experience will inherently be about fuzzy places or even multiple places rather than a point on a map.

The 10,496 locations are not uniformly distributed around the world, and in some places (e.g., in the London area), several location points exist in close proximity to one another. For the purpose of analysis, we created a one-quarter-degree grid over the Earth and aggregated all locations within a single grid square. A new point location was specified as the centroid of the grid square. Although a grid of that size consists of over a million cells, only 7,227 actually had associated entries. As an alternative method, the locations were also mapped to the United Nations Global Administrative Unit Layers (GAUL), a product of the Food and Agricultural Organization (FAO). That representation has three geographic layers mapped to county, state, and country represented in shapefile format, which allowed us to aggregate entries based on political boundaries.

12.4 Modeling Places from LDA Results

Given a set of georeferenced documents, D, the first step in our approach is to use LDA on the corpus to generate a set of topics, T. Following the LDA training, for each document, d, we have a location $\mathbf{x}^d = <x, y>$ and a T-dimensional vector, θ^d, specifying the multinomial distribution of topics for that document. Because the latitude and longitude specified for a georeferenced article is often an approximation of a vaguely defined region and there may be more than one article for a particular location, we need to relax the location. That is done by generating a fixed-sized grid over the Earth and averaging the topic distributions for each grid square, g_i, by finding the mean topic distribution vector, θ^{gi}, for all the documents spatially included within the square. Given the topic distributions for a set of grid squares, the next step is to spatially interpolate a continuous field representation for each topic.

Fig. 12.1 Sample latent topics from 200-topic LDA run

12.4.1 Topic Modeling Results

During our empirical tests, we ran several Gibbs sampling simulations on the dataset for 20, 50, 100, 200, 300, and 400 topics. For the selection of α, β, and topic number parameter values, we followed suggestions presented in Griffiths and Steyvers (2004). We kept the β value constant at 0.1 and the α value at 50/# of topics.

The topics that are discovered using LDA are often represented as an ordered list of the most commonly generated words for that topic. While these lists do rank the words in a topic from the most commonly generated word on down, they ignore the relative importance of the words. For example, in one topic, the top-ranked word might have probability 0.08 and the next most common word 0.01, which means the first word is eight times as commonly generated as the second. In other topics, the first and second word might have very close probabilities. We found that a word cloud visualization that shows the words in relative sizes is more illustrative and will therefore use that method in lieu of lists.

After observing the results, we found that the topics tend to fall into four broad categories: activity topics, feature topics, locality topics, and miscellaneous topics. Activities and features are distributions of words related to things to do and see, respectively. Locality topics consist of words that are associated with a specific geographic location. Miscellaneous topics are ones that do not appear to have any special relationship to traveling per se but nevertheless reflect semantic structures in the language. Many topics fall into more than one of these categories, which are fuzzy. Figure 12.1 shows some sample topics from a 200-topic run.

12.4.2 Adding Location Information

A trained LDA model results in a topic mixture for each blog entry. The LDA model does not include any spatial or temporal information as a parameter, so we do a post hoc analysis of the topic strengths for entries associated with specific locations. We propose that by combining the topic mixtures for all the entries in a location, an aggregate picture of that location's sense of place can be drawn. The best technique for aggregating the topic mixtures is not immediately apparent, however. The blog entries are not evenly distributed over the locations, and as a result, there are some locations with many more entries than at other locations. For our examples in this and the following section, we will identify the location of places with the one-quarter-degree grid squares as described in the previous section. One simple method of calculating a location topic distribution is to take the average θ values for each topic at the location. Let M be the number of entries for a location, θ_{ij} be the value for the ith topic of the jth entry, and L_θ be the location topic distribution:

$$L_{\theta_i} = \frac{\sum_{j=1}^{M} \theta_{ij}}{M}$$

The location topic distributions can then be used to calculate the similarity between places using the relative entropy measures described in Sect. 12.2 as a semantic distance measure. In order to get a similarity value [0.1], we define the similarity of places as an exponentially decaying function of this distance measure: $e^{-D_{KL}}$ (in this case using the asymmetric relative entropy). Alternately, one could use a linear or Gaussian decay function (Shepard 1987). Figure 12.2 maps out the similarity of places to Santa Barbara, CA. The results show a clear example of Tobler's first law of geography that generally speaking near places are more similar than far places (Tobler 1970), but it also captures that Santa Barbara shares some themes with some far places as well (such as urban areas in New York, for example).

This method treats locations with one or two entries as equal to locations with hundreds of entries. Depending on the goal of the analysis, this may or may not be problematic because locations with many entries (such as London) will tend to have a much smoother topic distribution due to the averaging. Intuitively, this makes sense because it is characteristic of global cities such as London that they are extremely heterogeneous and should reflect many different perspectives and themes. However, it is possible to normalize the distributions using the entropy of L_θ as the normalizing parameter.

An entropy measure, L_e, can be used to determine the degree to which a location topic distribution is about a few topics or many topics:

$$L_e = -\sum_{i=1}^{n} p\left(L_{\theta_i}\right) \log p\left(L_{\theta_i}\right)$$

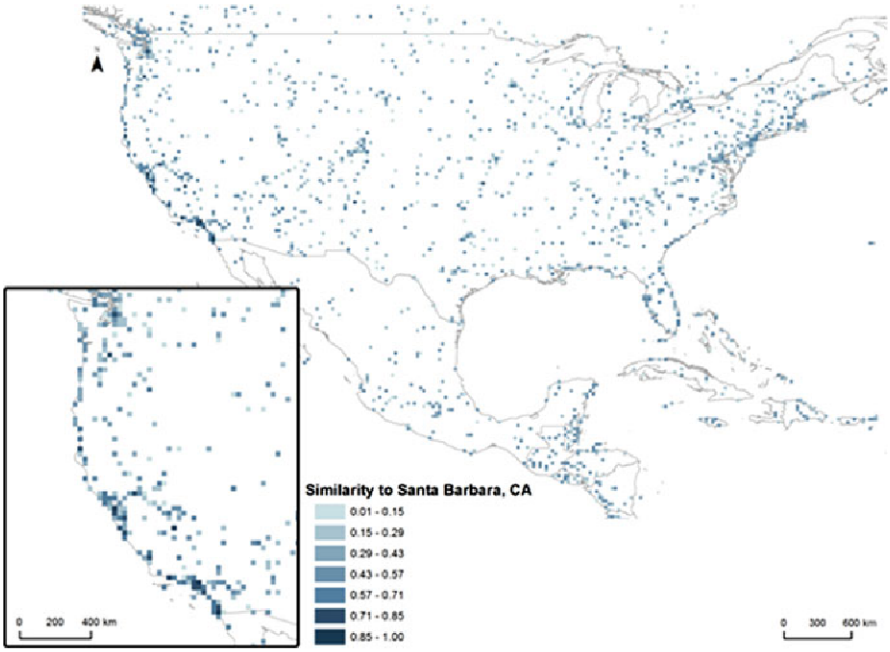

Fig. 12.2 Similarity to Santa Barbara, CA, based on relative entropy measures

The lower the entropy, the fewer the number of topics being written about for that location. Figure 12.3 shows the entropy of all places around Australia where we have five or more entries. This map illustrates that urban areas tend to have more diversity in topics being discussed than rural areas. Presumably, given that we are looking at travel blog entries, this is because certain rural locations are visited by people to undertake specific types of activities. Also, there is less heterogeneity in terms of the human geographic features in those places.

We propose that if an individual topic probability is prominent despite high overall entropy of the topic distribution for a location, it should be considered comparatively more important than an equivalent topic probability in a location with very low entropy. Let L_e be the location entropy, N be the number of topics, and γ be a scalar; the strength s_i of topic i is defined as

$$s_i = \begin{cases} \gamma L_e \theta_i, \ \theta_i > \dfrac{1}{N} \\ \theta_i, \ \theta_i \geq \dfrac{1}{N} \end{cases}.$$

The conditional is due to the fact that we only want to increase the prominence of topics that already have a probability greater than or equal to $1/N$.

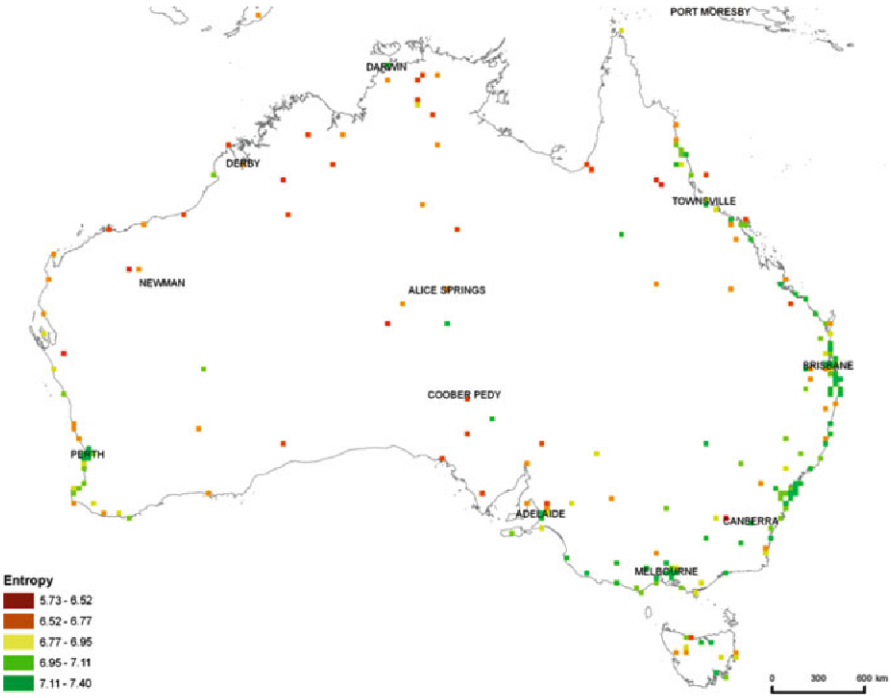

Fig. 12.3 Entropy of places in Australia

12.5 Constructing Regions of Thematic Distinction

Using the s_i as an indicator of the salience of topic i at a specific point location, we can use spatial statistics to generate a field representation of the topic's relevance around the world. Polygonal regions indicating where the topic is most relevant can be constructed by calculating a contour from the field representation based on a threshold value. In our examples, we identify topic strengths for the one-quarter-degree grid square centroids as described in Sect. 12.3. The topic strength at a point is treated as a point count input into an Epanechnikov kernel density estimation function (de Smith et al. 2007).

Figures 12.4 and 12.5 show results of mapping two topics using this method with contour lines at topic strength equal to 0.01.[2] The *wine* topic is shown for Europe. The *temple* topic illustrates the regionality of some topics – temple features are found much more often in South and East Asia, and this is reflected in what people write about in those places. Mapping out topics in this way is a two-way street. Not only does it provide a mechanism for exploratory analysis to find out

[2] It should be emphasized that the topic regions for the exemplars in this chapter reflect strong biases from the fact that original data are travel blog entries.

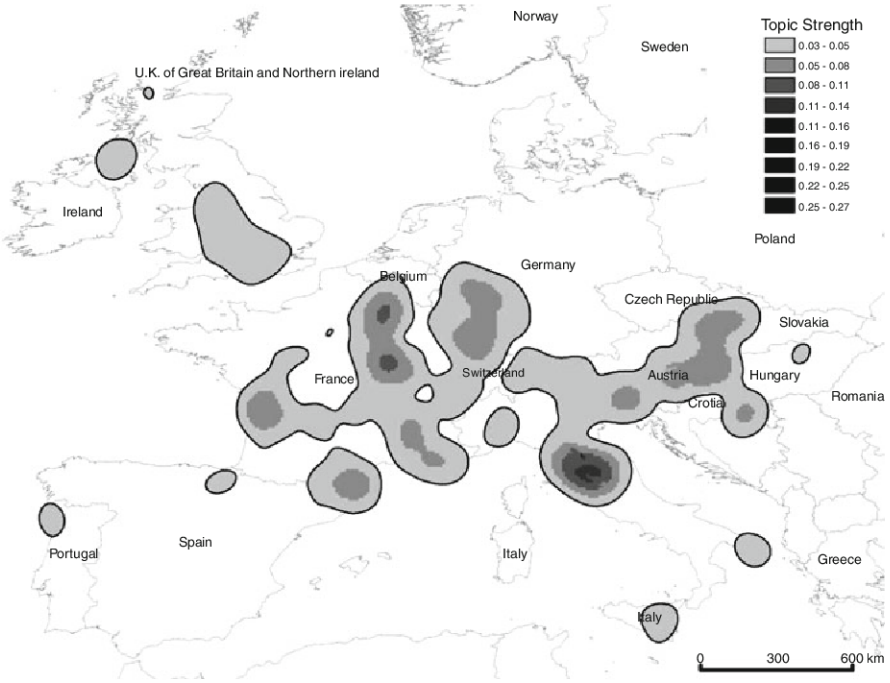

Fig. 12.4 *Wine* topic strengths

where certain topics are being mentioned but also in some cases the visualization can help in the interpretation of the meaning of the topic. For example, in our 400-topic simulation, two distinct topics (nos. 275 and 384) were generated that had *war* as the top-ranked word (see Fig. 12.6). Topic 275 appears to refer to war history, whereas topic 384 appears to be more about current war events. By mapping out the weights, we can begin to confirm those assumptions. Topic 384 is strong in current or recent hotspots (e.g., Iraq, Afghanistan, Sri Lanka), but topic 275 is not.

12.5.1 Visualizing Multiple Topics

Figure 12.7 shows a map of topics grouped into two themes. Topics related to human characteristics of place are shown in contrast to those related to physical characteristics. Through examination of the 200 topics, approximately 12 topics could be characterized as relating exclusively to the physical environment (e.g., mountains, rivers, beaches). Additionally, 13 topics are related specifically to human-constructed features (e.g., churches, cities, markets). Many other topics had both physical and human components. We aggregated the "physical" and "human" topics into two kernel density estimations created from the quarter-degree point values with a 1.5° radius.

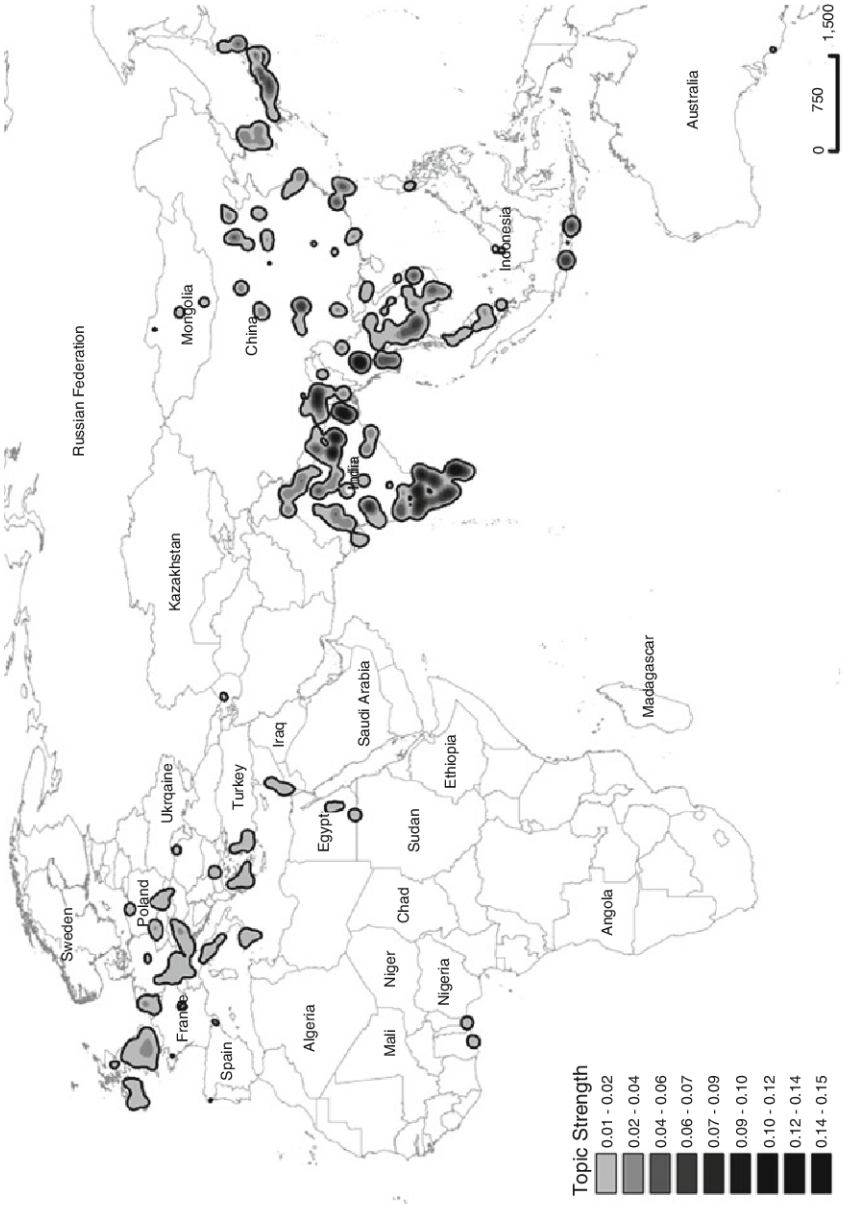

Fig. 12.5 *Temple* topic strengths

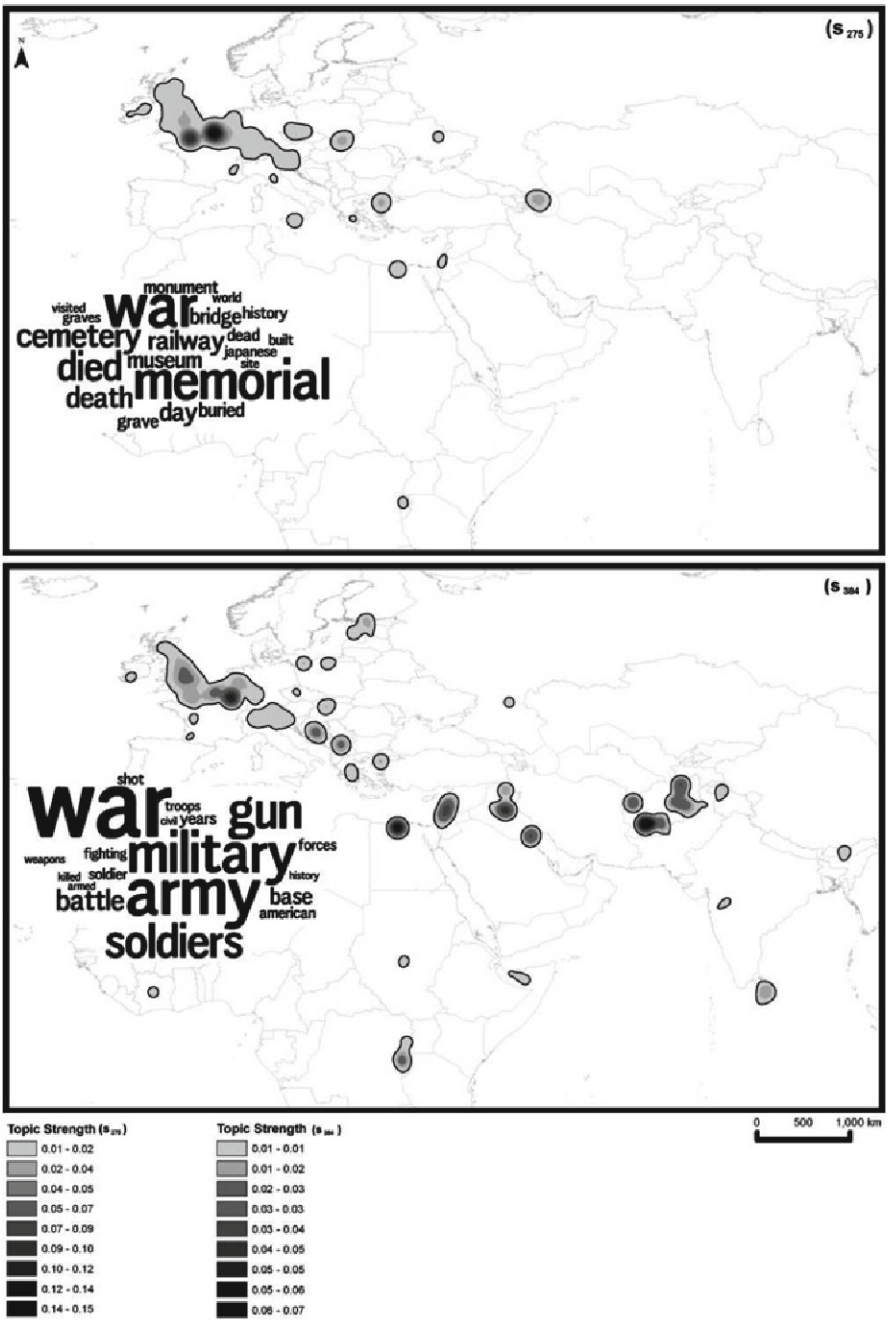

Fig. 12.6 Mapping topics 275 and 384

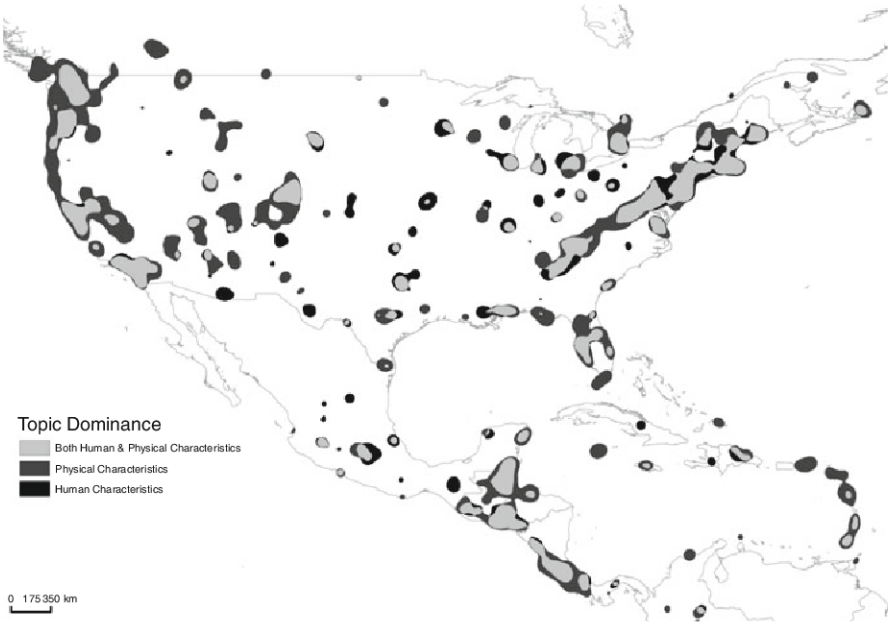

Fig. 12.7 Comparing regions based on strength of human vs. physical geography topics

The polygons displayed on the map show regions with a kernel density theme value above its mean. For example, the original physical characteristics theme values ranged from approximately 0.203–2.723 with a mean value of 0.457. The polygons shown on the map only contain values above 0.457.

12.5.2 Measuring the Localization of a Topic

As we have mentioned earlier, some of the topics generated by LDA have place names as top words and very clearly correspond to one specific region of the Earth. And as the maps in the previous sections show, there are some topics that show up in more than one part of the world but nevertheless are not evenly distributed everywhere (where *everywhere* in this context means all the locations for which we have blog entries). We would like to be able to measure the degree to which a topic is written about in one area or many or everywhere.

The topic strengths $S_i = \{s_{i_1}, s_{i_2}, \ldots, s_{i_n}\}$ for topic i at a set of n locations (e.g., the 7,277 grid cells with entries) can be interpreted as a kind of categorical probability distribution. The *localization* of topic i can then be evaluated as an inverse function of the entropy measure on that distribution. However, before localization can be

evaluated in this manner, the topic strengths must be normalized so that they sum to 1. Let k be a location; the "probability" of k for topic i is

$$p_k = \frac{s_{i_k}}{\sum_{j=1}^{n} s_{i_j}}.$$

Localization of topic i is then defined as follows:

$$\Lambda_i = \gamma e^{\sum_{k=1}^{n} p_k \log p_k},$$

where γ is a constant scalar value.

12.6 Temporal Analysis

In their analysis of scientific topics, Griffiths and Steyvers (2004) showed how a post hoc analysis of the linear trend line of the mean θ value for an LDA topic could be used to infer its "hotness" or "coldness." A similar technique can be used to identify the dynamic change in topics being written about in travel blog entries. To illustrate, we present some results from a 400-topic simulation that was run for 1,000 iterations. The mean theta value for each topic per day was calculated from all entries with at least 100 words that were written between January 1, 2006 and August 31, 2010.

Figures 12.8 and 12.9 show how a *Chinese* locality topic is trending upward over this time period while a *Japanese* locality topic is trending downward. The increase in variance of the point values in 2010 can be explained by the fact that fewer people were blogging on travelblog in 2010 than in previous years. While a linear fit is useful for revealing the overall drift in a topic's popularity, nonlinear fits can illuminate periodic patterns in a topic's popularity. For example, many of the LDA topics show seasonal fluctuations. A *festival* topic peaks during February and March when many festivals (e.g., Carnival and Mardi Gras) happen around the world (see Fig. 12.10). The strength of other topics coincides with specific types of events such as natural disasters. For example, topic 387 shown in Fig. 12.11 peaks after the May 2008 Sichuan earthquake. While it would require a much more in-depth analysis to verify, it is conceivable that some topics act as leading indicators for some types of events, especially ones that are socially constructed.

Combined with geographic information, this kind of analysis shows how some topics trend differently in different locations. A closer look at topic 387 shows that it, in fact, peaks at two different times in 2008 depending on the location of the entry. In China, predictably, it peaks in May 2008 after the Sichuan earthquake. However, in both the United States and India, it peaks when the Mumbai terrorist attacks occurred in October 2008. A review of the entries written at these places and times verifies that these events were referenced using words from the topic 387 distribution.

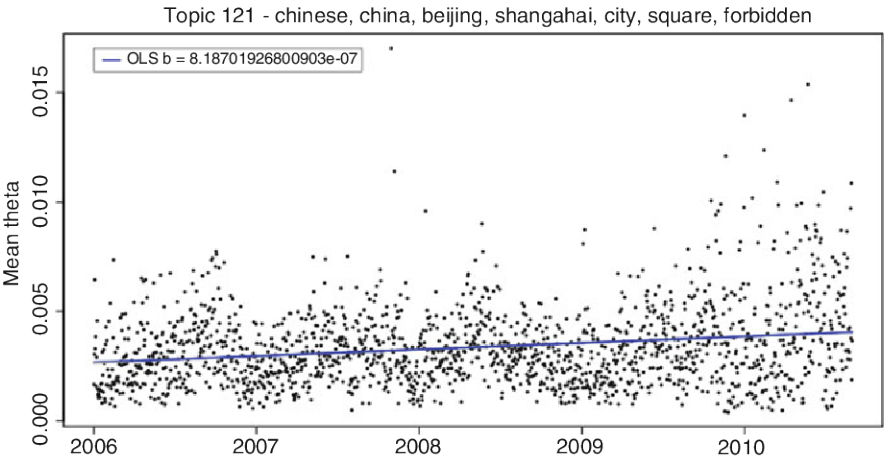

Fig. 12.8 *Chinese* topic trending upward

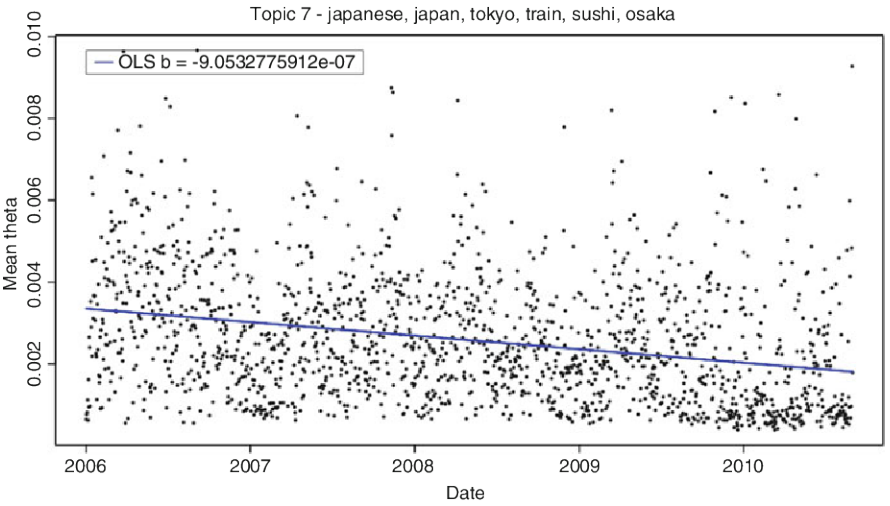

Fig. 12.9 *Japanese* topic trending downward

12.7 Summary and Conclusions

Much of the VGI available on the Web comes in the form of textual descriptions. In this chapter, we presented ways of using topic modeling to identify a place's unique mixture of characteristics directly from natural language observations. These methods allow us to calculate the similarity of places, map out places of thematic

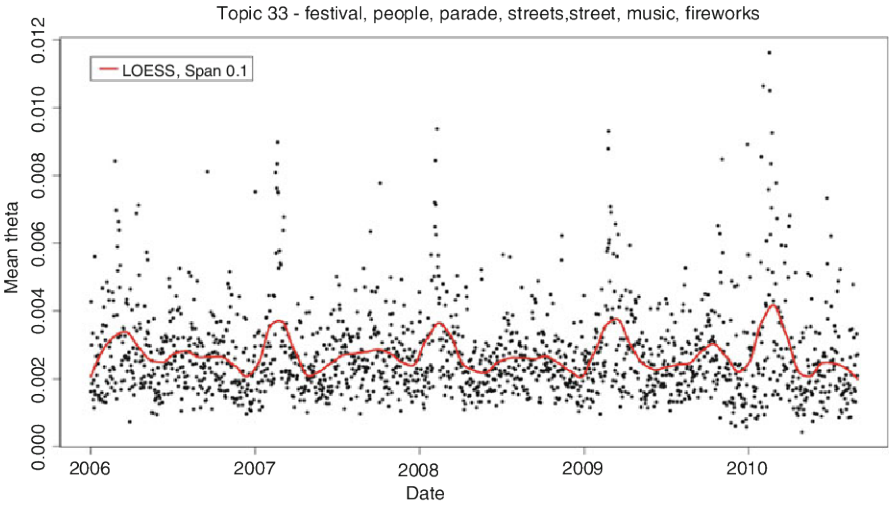

Fig. 12.10 *Festival* topic peaks in February/March

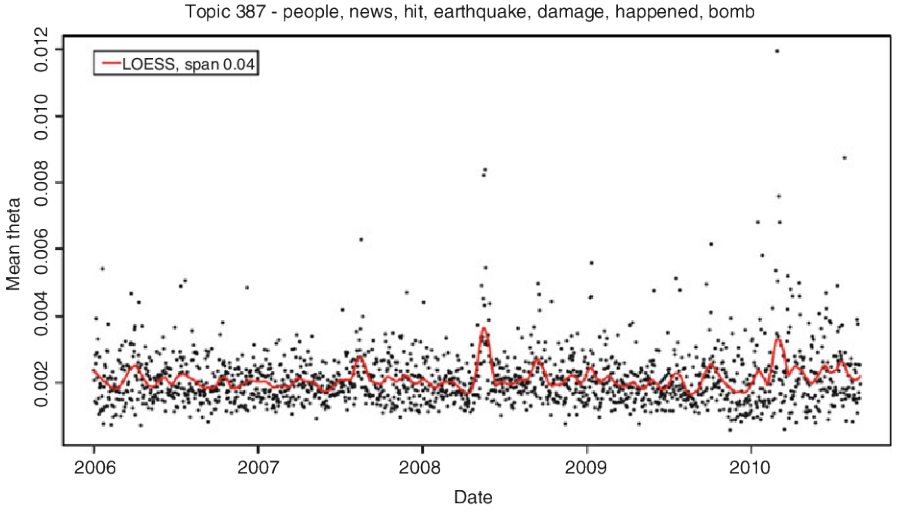

Fig. 12.11 Topic 387 strength corresponds to specific events

distinction, measure the degree to which certain themes are local or global, and evaluate thematic change over time. These results open up new opportunities for understanding the makeup and dynamism of places as described from individual experiences. A future task is to examine how these operationalized representations of place can be incorporated into a more robust framework that affords more in-depth reasoning about place, including a context-dependent similarity measure.

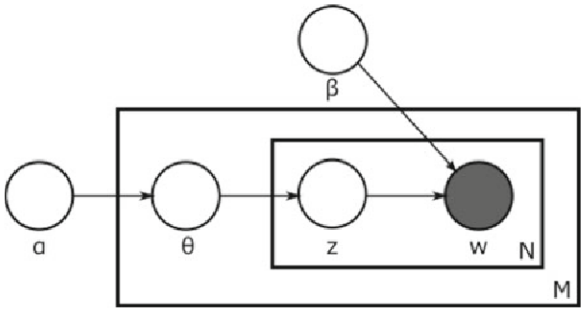

Fig. 12.12 Latent Dirichlet allocation model plate notation

The issue of granularity remains a problem in terms of locating the documents in space, and further work needs to be done to incorporate a granularity metric into the representations generated using our methods.

Acknowledgments The authors wish to thank Mike Goodchild and two anonymous reviewers for their valuable comments.

12.8 Appendix A: Latent Dirichlet Allocation Model

This appendix describes the generative model for LDA (Blei et al. 2003). Let α be the Dirichlet hyperparameter of the per-document topic distributions, β be the Dirichlet hyperparameter for the per-topic word distributions, θ_i be the multinomial topic distribution for document i, z_{ij} be the topic for the jth word in document i, and w_{ij} be the jth word. The generative model for LDA is then defined as follows:

- Choose θ_i proportional to Dirichlet(α), where $i \in \{1, \ldots, M\}$.
- For each of the words w_{ij}, where $j \in \{1, \ldots, N\}$:
  - Choose a topic $z_{i,j}$ proportional to multinomial(θ_i).
  - Choose a word $w_{i,j}$ proportional to multinomial($\beta z_{i,j}$).

Figure 12.12 shows the plate notation representation of the LDA model. Plate notation is a shorthand representation of graphical probabilistic models that have many repeating variables. Each circle represents a variable in the model and the number in the lower right corner indicates the number of times the variable is repeated. For example, θ is repeated M times in the model shown. The shaded variable w is the only observed variable (i.e., the words in the documents).

References

Agnew, J. (1987). *The United States in the world economy*. Cambridge, MA: Cambridge University Press.

Bishop, C. (2006). *Pattern recognition and machine learning* (Vol. 4). New York: Springer.

Blei, D. M., & Lafferty, J. D. (2006). Correlated topic models. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in neural information processing systems (NIPS) 18* (pp. 147–154). Cambridge, MA: MIT Press.

Blei, D. M., & Lafferty, J. D. (2009). Topic models. In A. N. Srivastava & M. Sahami (Eds.), *Text mining: Classification, clustering, and applications* (pp. 71–94). Boca Raton: CRC Press.

Blei, D., & McAuliffe, J. (2008). Supervised topic models. In J. C. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems (NIPS) 20* (pp. 121–128). Cambridge, MA: MIT Press.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Chang, J., & Blei, D. (2009). Relational topic models for document networks. In D. van Dyk & M. Welling (Eds.), *Proceedings of the 12th international conference on artificial intelligence and statistics (AISTATS)* (pp. 81–88). Clearwater Beach: Journal of Machine Learning Research.

Cresswell, T. (2004). *Place: A short introduction*. Malden: Blackwell Publishing Ltd.

de Smith, M., Goodchild, M., & Longley, P. (2007). *Geospatial analysis: A comprehensive guide to principles, techniques and software tools* (2nd ed.). Leicester: Winchelsea Press.

Entrikin, N. (1991). *The betweenness of place: Toward a geography of modernity*. Baltimore: Johns Hopkins University Press.

Goodchild, M. F. (2009). What problem? Spatial autocorrelation and geographic information science. *Geographical Analysis, 41*(4), 411–417.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(Suppl. 1), 5228–5235.

Hao, Q., Cai, R., Wang, C., Xiao, R., Yang, J. M., Pang, Y., & Zhang, L. (2010). Equip tourists with knowledge mined from travelogues. In M. Rappa, P. Jones, J. Freire, & S. Chakrabarti (Eds.), *Proceedings of the 19th international conference on world wide web (WWW'10)* (pp. 401–410). New York: ACM Press.

Jorgensen, B. S., & Stedman, R. C. (2006). A comparative analysis of predictors of sense of place dimensions: Attachment to, dependence on, and identification with lakeshore properties. *Journal of Environmental Management, 79*, 316–327.

Li, W., & McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. In *ICML'06: Proceedings of the 23rd international conference on machine learning* (pp. 577–584). New York: ACM Press.

McCallum, A. (2002). *MALLET:* A machine learning for language toolkit. http://mallet.cs.umass.edu. Accessed October 8, 2011.

Mei, Q., Liu, C., Su, H., & Zhai, C. (2006). A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In L. Carr, D. D. Roure, A. Iyengar, C. A. Goble, & M. Dahlin (Eds.), *Proceedings of the 15th international conference on world wide web* (pp. 533–542). New York: ACM Press.

Serdyukov, P., Murdock, V., & van Zwol, R. (2009). Placing flickr photos on a map. In J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, & J. Zobel (Eds.), *Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval* (pp. 484–491). New York: ACM Press.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*(4820), 1317–1323.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. In T. Landauer, D. Mcnamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 424–440). Hillsdale: Lawrence Erlbaum Associates.

Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004). Probabilistic author-topic models for information discovery. In *KDD'04: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 306–315). New York: ACM Press.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association, 101*, 1–30.

Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography, 46*(2), 234–240.

Tuan, Y. F. (1977). *Space and place: The perspective of experience*. Minneapolis: The Regents of the University of Minnesota.

Wang, C., Wang, J., Xie, X., & Ma, W. Y. (2007). Mining geographic knowledge using location aware topic model. In R. Purves & C. Jones (Eds.), *Proceedings of the 4th ACM workshop on geographic information retrieval* (pp. 65–70). New York: ACM Press.

Winter, S., Kuhn, W., & Krüger, A. (2009). Guest editorial: Does place have a place in geographic information science? *Spatial Cognition and Computation, 9*, 171–173.