On the Geo-Indicativeness of non-Georeferenced Text

Benjamin Adams[†] and **Krzysztof Janowicz**[⋆]
[†]Department of Computer Science and [⋆]Department of Geography
University of California, Santa Barbara
Santa Barbara, CA, USA 93010
[†]badams@cs.ucsb.edu and [⋆]jano@geog.ucsb.edu

Abstract

Geographic location is a key component for information retrieval on the Web, recommendation systems in mobile computing and social networks, and place-based integration on the Linked Data cloud. Previous work has addressed how to estimate locations by named entity recognition, from images, and via structured data. In this paper, we estimate geographic regions from unstructured, non geo-referenced text by computing a probability distribution over the Earth's surface. Our methodology combines natural language processing, geostatistics, and a data-driven bottom-up semantics. We illustrate its potential for mapping geographic regions from non geo-referenced text.

Introduction

Information retrieval and recommendation systems rely on explicit or inferred contextual information to better understand the user's needs. Location is among the most fundamental contexts; it is assumed that most data comes with some sort of spatial or geographic reference. While it is not surprising that mobile applications rely on location information, e.g., via GPS, location is also key to search on the Web and all major search engines and social networks mine location information that is not explicitly encoded in the user's query, e.g., via IP addresses. These references range from topological and metric relations, such as adjacency or nearness; references to types of features, such as *city* or *restaurant*; geographic names, e.g., Los Angeles; and spatial footprints, e.g., polygons, expressed in a spatial reference system. What makes using this information difficult is its vague nature. People and the data they create refer to *places* and not to explicit, well-defined *locations*. For instance, people may refer to vernacular names such as *American Riviera* or vague feature types such as *downtown* without clear, administrative borders.

Consequently, methods for estimating geographic location have been studied in different fields ranging from data mining to geographic information science to the social sciences. Different types of raw data have been used. While the majority of work includes named entity recognition, others exploit low-level image analysis to compute a probability surface for non-geotagged photos published on the Web (Hays and Efros 2008).

In this work, we propose to estimate the geographic regions to which an unstructured, non geo-referenced text likely refers. Our hypothesis is that natural language expressions are geo-indicative even without explicit reference to place names. For instance the set of terms {*manufacturing, traffic, income, skyline, government, poverty, employment*} contained in a textual description most likely refers to a larger city, while {*park, hike, rock, flow, water, above, view*} is rather associated with travel reports in national parks.

The remainder of the paper is structured as follows. First, we introduce background information relevant for the understanding of our work. Next, we present the estimation methodology, including the involved components, data sources, and measures, and illustrate our approach by showing examples. We then present experimental results for the region estimation and point out further research directions.

Background

In this section we provide background information on topic modeling and kernel density estimation, and provide a comparison to related work.

Topic modeling

Probabilistic topic modeling is a popular set of algorithms for identifying the latent topics in a corpus of documents. The most popular method, Latent Dirichlet allocation (LDA), is a generative model that models each document as the result of running a random process on a probabilistic graphical model (Blei, Ng, and Jordan 2003). Each document is modeled as a mixture of topics. The distribution of topics over documents and words over topics are assumed to follow the Dirichlet distribution for a given (often symmetric) hyperparameter vector. Each word in a document is generated by first selecting a topic given the topic distribution for that document followed by selecting the word from the topic. Given this model the challenge is to infer the topics and topic distributions that would have produced the observed corpus. Because it is a large probabilistic graphical model it is not feasible to solve

this Bayesian inference directly, so a number of approximation methods have been developed using variational bayes, expectation maximization, and markov chain Monte Carlo simulations (Blei, Ng, and Jordan 2003; Griffiths and Steyvers 2004).

Kernel Density Estimation

Kernel Density Estimation (KDE) is a frequently used spatial information analysis that estimates a local density surface of features in a certain range around those features (Silverman 1986). The probability density function f is estimated as given in equation 1 where h is the bandwidth and $K(\cdot)$ is some kernel function.

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big), \tag{1}$$

$$K(u) = \frac{3}{4}(1 - u^2)\, \mathbf{1}_{\{|u| \leq 1\}}, \tag{2}$$

We used the the Epanechnikov function (equation 2) but other distributions can be used as well. The bandwidth h can be fixed or variable and can be estimated.

Geographic Knowledge from Text

In recent years there has been increased interest in applying computational methods to better extract geographic knowledge from heterogeneous and unstructured data. Much work has been done on extracting facts and relations about named entities including named places from text using natural language processing, e.g., (Etzioni et al. 2005), and a number of web services now exist designed for that purpose. Integrating heterogeneous information for geographic sensemaking remains an active research area (Tomaszewski et al. 2011).

In addition, extensions of LDA have been developed for training based on labels, which could be used to identify topics associated with place names (Ramage et al. 2009). Another extension, the relational topic model, has been used to relate documents in a network of adjacent geographic locations (Chang and Blei 2009). Our approach differs from the previous, because we use spatial statistics post-hoc rather than conditioning topics based on a location tag or network structure. Performing this analysis post-hoc allows us to incorporate spatial dependence and generate thematic regions based on the spatial proximity of documents.

Named entity recognition can be used to georeference a document by simply matching the document to the coordinates or set of coordinates associated with the named places found in the document. This method is used by a number of web services, e.g., Yahoo Placemaker[1]. (Wing and Baldridge 2011) have proposed to predict geolocation from text using a geodesic grid, but our work differs in important ways. First, this method predicts the location of a document that has no place names in it while their methods are explicitly tailored to incorporate such knowledge. Second, our method derives a topic mixture from a weighted combination of the documents within a grid square. In contrast, the approach described in (Wing and Baldridge 2011) treats all text geolocated within a grid square as part of a single document and uses a Näive Bayes approach for training, disallowing documents that have different salience weights. Finally, we use kernel density estimation to identify contiguous regions that are thematically differentiable from other regions.

[1] `http://developer.yahoo.com/geo/placemaker/`

Estimation Methodology

To estimate the location to which a natural language description refers, we need to characterize how different textual themes are distributed over the Earth's surface. This requires a probability surface for each theme trained from a corpus of geo-referenced documents. There are many corpora on the Web that are geo-referenced to differing levels of granularity, which allows us to examine how and to what degree topics in different corpora are geo-indicative. For this study we chose two corpora from different domains: Wikipedia articles and travel blog entries. The resulting probability surfaces differ with the used corpora; it is critical to consider the data source before performing any additional analysis.

Since we are interested in estimating geographic regions based on non geo-referenced, unstructured text, we pre-process our training corpora by identifying and removing references to named places. We do so by using Yahoo's Placemaker Web service. To maintain control over the used language we apply a language detection script to identify and remove all non-English documents. Finally, we remove all words from a standard English stop word list and stem the remaining words as is common practice in information retrieval tasks.

Formally, we define the corpus D as a set of documents and let P be a set of geo-referenced points. Each document is interpreted as bag-of-words and used to train topics using LDA. The LDA training results in a set of T topics and a $\mid T \mid$-dimensional probability vector for each document, θ_d.

Probability Surface and Regions for Topics

Following the LDA training, we estimate a probability surface on the Earth for each topic. In order to do this, we need to spatially interpolate a field representation for each topic, z, from a set of observation data, namely each document's θ_{dz} value and spatial representation (i.e., points or polygons). Empirically, we found that traditional spatial interpolation methods, such as inverse distance weighting, do not produce interpretable results for our data sets because there is a very high variability for an individual topic's values over a short distance. Looking at Wikipedia for example, this is intuitive because an article about a baseball stadium can be located next to an article about a community center and these two articles will likely have very different

topic profiles. We interpret a document's topic value as a point count describing how often that topic is observed at the document's location. Then we use KDE to estimate a probability surface for the topic based on those point counts. However, since KDE has an additive effect we need to first normalize the densities of articles, so that places with many documents (e.g., London in Wikipedia) do not result in high probability peaks for all topics.

We create a grid, G, over the Earth set to a fixed width and height based on decimal degrees. Due to the Earth's curvature a simple grid based on latitude and longitude will not result in equal sized grid squares and depending on the spatial distribution of the documents this may be a problem that can be solved using an alternate equal area grid instead. However, we did not find any discernible problem with this approach for region generation because adjacent grids are very close to equal size. Moreover, our corpora only contain a few geocoded articles to very northern or southern latitudes where this distortion is more apparent.

We then calculate the spatial intersection of documents with grid squares, so that for each grid square, g, we can calculate a vector of topic values, $\theta_{\mathbf{g}}$. We set g_z, the z topic value for g, equal to a weighted average of the θ_{dz} values of the intersecting documents, where the weight for each document is w_d. For each topic we create a point set corresponding to the centroids of all grids with $g_z > 0$ and perform KDE to generate a field representation for each topic stored as a raster file. The default bandwidth of the kernels is set to twice the grid square width allowing for region generation when high topic values are found in adjacent grids. The resulting raster file can be further processed to identify topical regions by drawing a contour polygons; see figure 1.

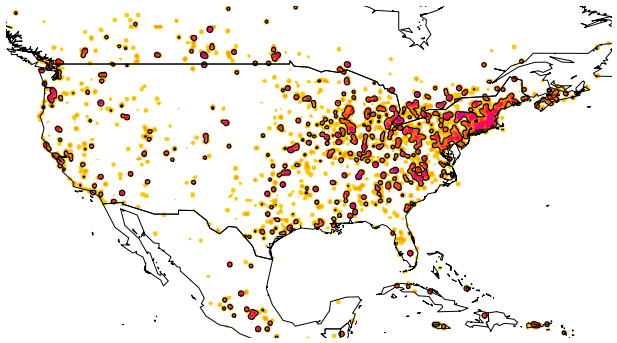

Figure 1: Probability surface for a topic from the Wikipedia model with regions defined by a threshold contour. The top words include *industi, factori, compani, manufactur, product, work, plant, employ, worker, busi, build, make, oper, larg, process.*

Geographic Profile from Text

Following the creation of KDE rasters for each topic we use the standard spatial analytic approach of building a weighted raster overlay to generate probability surfaces for arbitrary mixtures of topics. Though the initial topic modeling inferencing and KDE calculation are computationally intensive operations, they can be performed offline. In contrast, the calculation of a weighted sum is very fast and can be performed online. To identify the topic distribution for the text the trained LDA model is used to infer a topic distribution for the new document (Griffiths and Steyvers 2004).

We can quantify how geo-indicative an individual topic, z, is by measuring whether the topic is rather evenly distributed over the Earth or concentrated in a few places. We do this by normalizing the topic values for all g in G so they sum to 1. We treat them as a multinomial probability distribution and compute the entropy, S_z. Topics with lower entropy are interpreted as more geo-indicative (see table 1).

Entropy	Topic words
Sample from Wikipedia 500 topics, highly geo-indicative	
6.52	t479: council, local, parish, district, servic, includ, villag
6.78	t265: bus, interchang, servic, termin, station, loop, road
6.87	t44: popul, people, sector, year, municip, swiss, area
Sample from Wikipedia 500 topics, non geo-indicative	
12.26	t314: year, time, mani, dure, earli, centuri, larg
12.21	t76: onli, mani, howev, main, larg, small, area
12.12	t120: south, north, east, west, area, border, part
Sample from travelblog 1000 topics, highly geo-indicative	
6.01	t279: nepali, squar, durbar, templ, thamel, nepales, stupa
6.39	t857: everest, base, day, jacinta, trek, namch, altitud
6.51	t372: gili, island, boat, air, beach, trawangan, lombok
Sample from travelblog 1000 topics, non geo-indicative	
11.47	topic 211: morn, day, earli, befor, arriv, back, afternoon
11.40	topic 959: decid, found, head, back, find, didn, ourselv
11.33	topic 889: day, veri, good, place, great, made, amaz

Table 1: Selected high and low geo-indicative topics for the Wikipedia and travel blog corpora

The topic distribution of the document, θ, is then used as input to create a unique probability surface for the document calculated as a weighted raster overlay via map algebra. The surface is a weighted raster overlay of all topic rasters where the topic value is greater than $1/\mid T \mid$. In other words we want to only consider topics that have a greater than random chance for the document. The weight, ω_z, for topic z is a product of its geo-indicativeness, i.e., normalized inverse entropy, and the topic weight for the document. Using these weights the surface for the document is calculated as a weighted raster overlay via map algebra.

Figure 2 illustrates the geo-indicativeness of the topics from a sample geo-located Wikipedia article (Santa Barbara, CA) with all place names removed.

Estimating Location

To evaluate our method we used sets of 200 held-out documents from Wikipedia and the travel blog datasets. Figure 3 plots the percentage of documents against threshold distances to the estimated locations. The results improved when the best of the top-N (N=30) locations is used, showing that the true location is often in the top-N results. Considering that the texts have all identifying place names removed this result shows that

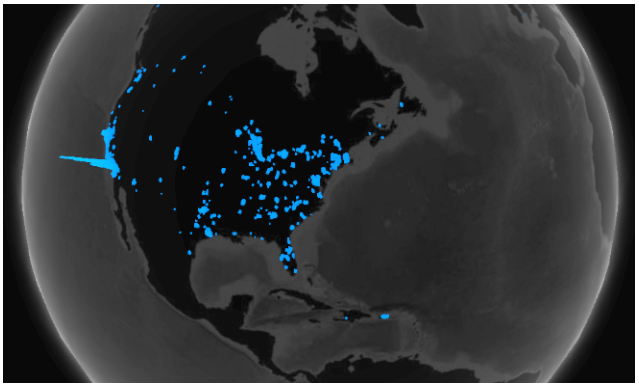

Figure 2: Mapping the weighted topic mixture for Wikipedia's *Santa Barbara, CA* article with all place names pruned before.

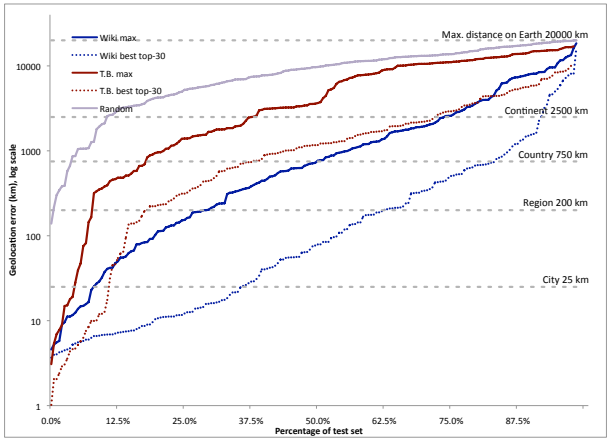

Figure 3: The distances between estimated location and actual location for each test document were sorted and plotted to show what percentage is successfully located within a given distance. Results for the maximum probability peak and the best of top-30 are shown. The maximum estimation error on the Earth using great circle distance is ≈ 20,004 km.

thematic content can be used to predict geographic location at a far better than random rate.

The prediction for Wikipedia articles is substantially better than for travel blogs, but this result is to be expected, since the topical content of travel blogs are less constrained. In addition, the mapping from location to the textual content is more tenuous for the travel blog entries. For more stylized text like Wikipedia articles the geo-location is very successful considering that all place names have been removed. For best of top-30, 75% of the test Wikipedia articles were located within 505 km and 50% were located within less than 80 km of the true location. In other words, the automatic matching of non geo-referenced plain text excluded more than 99% of the land surface of the earth. For the travel blogs 50% of the entries were predicted within less than 1169 km. While both approaches and datasets are not directly comparable, the presented results are in line with those reported for photo geolocation (Hays and Efros 2008, fig. 6). Note, however, that they use a top-120 and, surprisingly, their upper bound is the Earth's diameter (12,742 km) which appears to be an error.

Conclusion

In this work, we have shown how to estimate geographic regions corresponding to unstructured, non geo-referenced text. We proposed a methodology that learns LDA topics from geocoded Wikipedia and travel blogs as training corpora. Together with KDE, the resulting topics are used to estimate which geographic regions a new text is likely about by computing combined probability surfaces over the Earth's surfaces. This estimation is possible because the learned LDA topics are geo-indicative. In the future, we will further exploit these methods to infer the types of the described geographic features using semantic signatures.

References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Chang, J., and Blei, D. 2009. Relational topic models for document networks. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 81–88.

Etzioni, O.; Cafarella, M.; Downey, D.; Popescu, A.-M.; Shaked, T.; Soderland, S.; Weld, D. S.; and Yates, A. 2005. Unsupervised named-entity extraction from the web: an experimental study. *Artificial Intelligence* 165(1):91–134.

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

Hays, J., and Efros, A. A. 2008. im2gps: estimating geographic information from a single image. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society.

Ramage, D.; Hall, D.; Nallapati, R.; and Manning, C. D. 2009. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*, 248–256. ACL.

Silverman, B. W. 1986. *Density estimation: for statistics and data analysis. Monographs on Statistics and Applied Probability 26.*

Tomaszewski, B. M.; Blanford, J.; Ross, K.; Pezanowski, S.; and MacEachren, A. M. 2011. Supporting geographically-aware web document foraging and sensemaking. *Computers, Environment and Urban Systems* 35(3):192–207.

Wing, B., and Baldridge, J. 2011. Simple supervised document geolocation with geodesic grids. In *ACL*, 955–964. The Association for Computer Linguistics.