Geolocating Blogs From Their Textual Content

Clay Fink,[1] Christine Piatko,[1] James Mayfield,[1] Tim Finin,[2] and Justin Martineau[2]

[1]The Johns Hopkins University Applied Physics Laboratory, 11100 Johns Hopkins Road, Laurel, MD 20723

[2]The University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250

clayton.fink@jhuapl.edu

Abstract

Mashups showing the geographic location of the authors of social media content are popular. They generally depend on the authors reporting their own location. For blogs, automated geolocation strategies using IP address and domain name are not adequate for determining an author's location. Instead, we detail textual geolocation techniques suitable for tagging social media data, facilitating development of geographic mashups and spatial reasoning tools.

Introduction

Web sites such as Feedmap.net, GeoURL, and Twittervision allow people to associate their blog or "tweets" with a location and provide mashups that give a geographic view of Web content. These sites depend on content contributors providing their location explicitly. We have investigated whether it is possible to discover an author's location automatically.

Geolocating blogs using IP address and domain name is not always a viable strategy since most blogs are hosted by services like Blogspot, Wordpress, and Livejournal. During May and June 2008, we crawled approximately 800,000 blogs that pinged the weblogs.com ping site. Only 3% of them had unique IPs, whereas 82% were hosted on IPs with at least 100 other crawled blogs. Even if a blog is self-hosted, there are questions regarding the accuracy of IP-based and domain name-based geolocation techniques [1]. Thus, depending on these techniques alone is not sufficient.

Many bloggers supply their location as text on their blog homepage or on an "about me" page. Such text, however, is not guaranteed to be expressed in a standard format. HTML meta tags such as ICBM and geo.position allow the author to supply his or her position as latitude and longitude. In the crawl described above, we found that only 900 blogs out of 800,000 blogs had such tags. This suggests that these tags are not widely used.

We describe how to infer an author's location from textual mentions in their blog posts using techniques described in [2,3,4]. For example, we could infer that a blog post containing the strings *New York*, *Upper East Side*, *Central Park*, and *Gramercy Park* is about New York City. Similarly, we could infer that another post from the same blog containing the strings *Baltimore*, *Catonsville*, and *Camden Yards* is about the area around Baltimore, Maryland. Given enough mentions of location entities from a particular geographic area, one might confidently assert that that area is the geographic focus of the blog. To test this approach, we compared the extracted geographic foci of a set of ground truth blogs to their known locations.

Method

For each blog, we treat the home page and each post as a separate document. We apply three processing steps to each document: location entity recognition, the disambiguation of place names, and the determination of geographic focus.

We strip all HTML tags from a document's content and use a named entity recognizer [5] to extract location entity mentions from the text. Each entity name is matched against the GeoNames online gazetteer (http://www.geonames.org), producing a list of toponyms with associated latitude, longitude, and hierarchal administrative data (county, state, country, etc.).

We make multiple disambiguation passes over each document. The first pass disambiguates any name that has a toponym that is a continent, country, or first-level administrative area (e.g., a U.S. state), or a national capitol to that toponym. A second pass looks for cases where an ambiguous name is qualified by a disambiguated name. For example, if "Maryland" has already been disambiguated as the U.S. state of Maryland, then the text "Laurel, Maryland" leads us to disambiguate "Laurel" as Laurel, Maryland. The final pass disambiguates any remaining mentions to their most populous toponym. Once an entity name is disambiguated, we adopt the "one sense per discourse" paradigm and assume that any mention in any subsequently processed document for a given blog refers to that place.

To determine the geographic focus of a blog, we adapted the focus algorithm described in [2]. Using an OWL ontology that captures the structure of the toponyms stored in the GeoNames gazetteer, we created an ontology instance that captures the hierarchal relationships of all the disambiguated place names. For example, Baltimore, Maryland, would be found in the subhierarchy Baltimore/City of Baltimore/Maryland/United States. We assign each node representing a disambiguated location an initial score, the value of the score conferring some measure of the importance of the location toward inferring a geographic focus.

For example, initial scoring could be based on the confidence associated with the disambiguation. The remaining nodes in the hierarchy are given an initial score of zero. We then apply a scoring algorithm, following the work in [2], that propagates the scores up the hierarchy, starting at the leaf nodes, with the score decaying as the location becomes more general. Let n be a node at level $L \geq 1$ of the hierarchy, with $L = 0$ at the leaf nodes. Let I_n = initial score of node n. Let s_i be the accumulated score, or initial score if it is a leaf node, of child i of node n. Finally, let D be a decay constant where $0 < D < 1$. The accumulated score, s_n, of node n is

$$s_n = I_n + \sum s^{2}_{i} D^{L-1}$$

After applying this algorithm to the entire hierarchy graph, the higher scoring nodes - ignoring nodes for the globe and continents - will represent regions containing more disambiguated place names than those represented by lower scoring nodes.

Experiment

We collected approximately 1,000 English language blogs by authors who self-reported their location as the United States. The blogs were identified by crawling the weblogs.com ping server and searching for blogs with the HTML meta tags ICBM or geo.position. Additional blogs were taken from feedmap.net, where authors can register their locations. We retrieved posts for each blog using the Google Reader API going back as far as data was available. All blogs used were updated regularly (more than twice a month) and recently (since June 1, 2008), and we also screened out spam blogs [6]. The blogs were then checked manually to determine if the blogger's reported location was accurate. The location was modified if it did not match the author's actual location, which was determined by reading the content of some of their posts. Blogs for which we could not verify the location were not used.

We tested our algorithm against 500 blogs from our collection, using posts authored between January 1, 2005, and November 1, 2008. The scores for disambiguated place names in the hierarchy were initialized with an initial value of 0.5, and we used a decay constant of 0.8. For blogs where there was insufficient clustering of the nodes to cause the propagation of scores up the hierarchy, we ignored the result. To select the geographic focus of a blog, we traversed down the hierarchy, starting at the highest scoring node, and selected the subnode that was lowest in the hierarchy and had the highest accumulated score. A correct result was defined as being when the extracted geographic focus subsumed the blog's true location, or was within 100 miles of it. We had 295 matches out of 481 usable results for 61% accuracy. For the 295 matches, the average distance from the extracted location to the known location was 50.8 miles .

Conclusion

Our results suggest that for many blogs, the extracted geographic focus does indeed correspond to the author's location. Our future work will include improving the accuracy of this algorithm. A named entity recognizer trained on blog text might improve performance since the one used was trained on text from news stories. Better disambiguation strategies might also improve performance. Both improvements would provide a larger number of disambiguated place names for determining geographic focus. A hybrid approach that combines the use of IP and domain name geolocation, and any useful metadata, with our technique might also improve performance. Handling cases where multiple geographic foci are reported might boost our accuracy as well.

We will also investigate two hypotheses. The first is that we can learn to distinguish blogs for which the technique is effective from those where it is not. This will likely depend, for example, on blog attributes such as blog genre (e.g., diary vs. professional), topic, or number of authors. The second is that we can learn to classify toponym mentions that are likely to refer to locations near the author from those that are not, based on features in the surrounding text. For example, toponyms in sentences containing *I* and *we* may be more likely to provide evidence of the author's location at different points in time.

References

[1] Muir, J., van Oorschot, P. (2006). *Internet geolocation and evasion* (TR-06-05). School of Computer Science, Carleton University.

[2] Amitay, E., Har'El, N., Sivan, R., & Soffer, A. (2004). Web-a-where: Geotagging Web content. *Proc. of the 27th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 273-280.

[3] Smith, D., & Crane, G. (2001). Disambiguating geographic names in a historical digital library. *Proc. of the 5th European Conf. on Research and Advanced Technology for Digital Libraries*, 127-136.

[4] Zong, W., Wu, D., Sun, A., Lim, E. P., & Goh, D. H. (2005). On assigning place names to geography related Web pages. *Proc. of the 5th ACM+IEEE Joint Conf. on Digital Libraries*, 354-362.

[5] Finkel, J, Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics*, 363-370.

[6] Kolari P., Java A., Finin, T., Oates T., & Joshi, A. (2006), Detecting spam blogs: A machine learning approach. *Proc. of the 21st National Conf. on Artificial Intelligence*.