The Geolocation of Web Logs From Textual Clues

Clayton Fink, Christine Piatko, James Mayfield, Danielle Chou
The Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Road, Laurel, MD 20723

Tim Finin, Justin Martineau
Department of Computer Science
The University of Maryland, Baltimore County
1000 Hilltop Circle, Baltimore, MD 21250

Abstract— **Understanding the spatial distribution of people who author social media content is of growing interest for researchers and commerce. Blogging platforms depend on authors reporting their own location. However, not all authors report or reveal their location on their blog's home page. Automated geolocation strategies using IP address and domain name are not adequate for determining an author's location because most blogs are not self-hosted. In this paper we describe a method that uses the place name mentions in a blog to determine an author's location. We achieved an accuracy of 63% on a collection of 844 blogs with known locations.**

Keywords- social media; geolocation, named entity recognition;disambiguation

I. INTRODUCTION

As of March 2008, an estimated 184 million people were blogging worldwide [10], and there has been an understandable interest in the geospatial distribution of the authors of all this content. 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 social media content on the Web. These sites depend on content contributors providing their location explicitly. Automated geolocation strategies that can, with some precision, extract the geographic location of an author would be useful in these and other types of applications. In particular, applications that attempt to map the geospatial distribution of sentiment or opinion—in the aggregate—would benefit from automated geolocation because a very large amount of content will need to be processed, and it is not likely that an adequate portion of this content will have been geo-tagged manually.

Geolocating blogs using IP address and domain name is not always a viable strategy because most blogs are hosted by services such as 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, the accuracy of IP-based and domain name–based geolocation techniques is questionable [4]. Thus, depending on these techniques alone is not sufficient.

Many bloggers supply their location as text on their blog home page 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, however, we found that only 900 blogs out of 800,000 blogs had such tags, suggesting that these tags are currently not widely used by authors.

We have investigated using the textual content of blog posts to infer an author's location. Many bloggers will mention place names in posts. Given a single post, the locations mentioned may cluster in a particular geographic area, giving the post a unique geographic focus. For example, we could infer that a blog post containing the strings New York, Upper East Side, Central Park, and Gramercy Park has Manhattan as its geographic focus. Similarly, we could infer that another post from the same blog containing the strings Baltimore, Catonsville, and Camden Yards has the area around Baltimore, Maryland as its geographic focus. Given enough mentions of location entities from a particular geographic area across all of a blog's posts, one might assert that that area is the geographic focus of the blog.

It has been estimated that among people who blog, 64% blog about their personal life [10]. It is reasonable to assume that some will make mention of locations near where they live when describing their daily activities. We ask the question whether a blog's geographic focus, as extracted from textual clues across all posts, coincides with author's location. To test this approach, we compared the extracted geographic foci of a set of ground truth blogs with their known locations.

In the next section we discuss related work on disambiguating place name mentions and extracting geographic focus from document text. We next describe the method we used to process documents and extract geographic focus from place name mentions. Finally, we report our experimental results, discuss their implications, and point to areas of future research.

II. RELATED WORK

Our approach can be broken down into three subtasks: named entity recognition (NER) to extract location entity mentions from a document, the disambiguation of the extracted location entity mentions to points on the map, and determining the geographic focus of a document. Named entity recognition has been widely studied and is too broad a topic to review here, so we focus only on work relating to place name disambiguation and extracting the geographic focus of documents.

Common among all of the disambiguation strategies described is the use of a *gazetteer*: a collection of geographic locations that provides for each location a latitude and longitude, the location's classification (country, state, city, park, building, etc.), its population (if applicable), and information on the location's inclusion hierarchy (city, state, country and continent). For a given location entity mention extracted from text, a gazetteer can be used to find a list of toponyms that match the entity's name and represent candidate locations for that entity. Various algorithms have been described that use such candidate locations to disambiguate, or to ground, an extracted entity mention to a location on the map by assigning a specific toponym to it.

Smith and Crane [8] first looked at the local context of an entity mention and took into account cases where a location name is qualified by another location name (for example, "Laurel, Maryland"). They then looked at the document-level context of the mention and assumed that toponyms that are close together geographically can be used for disambiguating their corresponding mentions. The centroid of the positions of all toponyms for all mentions in a document was calculated, and all positions greater than two standard deviations from the centroid were dropped. This was repeated until a convergence criterion was met. They then processed the remaining toponyms in the sequential order of the mentions in the document, using a sliding window of four mentions. In each window, they scored the unresolved mentions on the basis of the distance of their toponyms to resolved mentions in the window, the calculated centroid of the document, and the relative importance of the toponym (i.e. country > city > park). The highest scoring toponym was selected.

Leidner et al. [5] also began by looking at the local context of a mention. They then used two "minimality heuristics." They first applied the assumption of "one referent per discourse," where all mentions of a given place name in a document are assumed to refer to the same location on the map. The second heuristic is the assumption of "spatial minimality," where the smallest geographic region that encloses a set of toponyms for all mentions in a document is assumed to give the correct assignment of toponyms to mentions.

Amitay et al. [1] processed each mention by first disambiguating on the basis of local context. They then applied a heuristic whereby the toponym with the largest population is assumed to be the correct assignment. Next they looked for a "disambiguating" context in the document for any unresolved mentions by looking for toponyms of other mentions in the document that were nearby or enclosed the toponym for the unresolved mention. The assumption of "one referent per discourse" was also applied as part of this method.

Zong et al. [9] again used local context to start off their disambiguation process. They then looked for mentions that had a unique match in the gazetteer. Using the already resolved mentions, they looked for linguistic patterns that implied a containment relationship or association between place name mentions.

A number of researchers are also using Wikipedia as a resource for disambiguating place names. Overell and Rüger [7], for example, crawled Wikipedia articles and built co-occurrence models for place names that can be used for disambiguation location entity mentions in text.

Both Amitay et al. and Zong et al. also extracted geographic focus from Web pages, both using tree-based algorithms that use the hierarchal data available in gazetteers. Amitay et al. populated a tree containing all disambiguated mentions along with all the locations in their inclusion hierarchy. They then scored the disambiguated mentions on the basis of the confidence associated with their disambiguation. A scoring algorithm was applied to the tree that gave a preference to nodes lower in the hierarchy with larger sub-trees. The highest scoring node was assumed to be the geographic focus of the document. They tested their algorithm on 20,000 Web pages with human-determined geographic foci and achieved a 92% accuracy at the country level, 30% accuracy at the city or state level, and 38% precise matches.

Zong et al. similarly built a tree based on the gazetteer hierarchy for disambiguated mentions. They used a scoring based on the frequency of each disambiguated mention in the document and an algorithm that takes into account the entropy in the scores of child place names. Unlike Amitay et al., who extracted the geographic focus of an entire Web page, Zong et al. extracted place name assignments for each segment of a Web page in its HTML DOM hierarchy. They achieved an accuracy of 66% on a set of Web pages with locations assigned by human subjects. A higher accuracy of 87% was obtained when erroneous disambiguation results involving nongeographic entities being resolved as geographic entities were dropped from the data.

III. METHOD

Our approach attempted to extract a geographic focus for a blog by aggregating location entity mentions across all of a blog's post. Post comments were excluded from this analysis because they usually contain contributions from third parties, and we were interested only in the author's contributions. We processed posts sequentially by post date and applied named entity recognition and location entity disambiguation to each post. We then used the disambiguated locations extracted from all posts to determine the geographic focus of the blog.

For each post, we stripped all HTML tags the content and used a named entity recognizer [2] to extract location entity mentions from the text. We used both location and organizational entities because entities with geospatial properties, such as universities and museums, were tagged as organizations by the NER we used. Each entity name was matched against the GeoNames online gazetteer (http://www.geo-names.org), producing a list of toponyms with associated latitude, longitude, and hierarchal administrative data (county, state, country, etc.). GeoNames contains 6.5 million unique geographic features, including 2.2 million populated places, and was populated with data from the U.S. National Geospatial-Intelligence Agency, the

U.S. Geological Survey Geographic Names Information System, and other international sources of geographic data. Although GeoNames provides a web service for accessing the gazetteer, we downloaded the data and populated a MySQL database for our work.

Prior to disambiguation, we dropped entities that had been tagged as more than one entity type in the text. For example, a person's first name, in isolation, may also be tagged as a location. We also dropped nongeographic entity mentions that were commonly misclassified as locations by the NER and were also in the gazetteer. For example, "Obama" and "Coca Cola" were on occasion tagged as locations. Obama is a city in Japan, and Coca Cola is a populated place in Panama and are both in the GeoNames gazetteer. We added such entity names to a set of "stop places" that was used to screen the entity names returned from the NER.

We used three simple heuristics for disambiguating place names and made a separate pass over a post for each technique. The first pass disambiguated any mention that was a continent, country, first-level administrative area (e.g., a U.S. state), or national capitol or was unique within the gazetteer to the applicable toponym. A second pass applied the "local context" heuristic and looked for cases where an ambiguous mention was qualified by an already disambiguated mention. For example, if "Maryland" had already been disambiguated as the U.S. state of Maryland, then the text "Laurel, Maryland" led us to disambiguate "Laurel" as Laurel, Maryland. The final pass disambiguated any remaining mentions to their most populous toponym. Once an entity name was disambiguated, we adopted the "one referent per discourse" paradigm as well and assumed that any mention in any subsequently processed post for a given blog referred to that place.

To determine the geographic focus of a blog, we adapted the focus algorithm described in [1]. Using an OWL ontology that captures the structure of the toponyms stored in the GeoNames gazetteer, we created an ontology instance that captured 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 assigned 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 were given an initial score of zero. We then applied a scoring algorithm, following the work in [1], 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 nonleaf node at level $L - 1$ of the hierarchy, with $L \geq 1$ and the leaf nodes at level 0. 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 will represent regions containing more disambiguated place names than those represented by lower scoring nodes.

IV. Experiment

We collected 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 were 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 [3]. 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 the 844 blogs in our collection that met our screening criteria, using posts authored between January 1, 2005, and April 24, 2009. 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 not sufficient clustering of the nodes to cause the propagation of scores up the hierarchy, we ignored the result. There were 25 blogs with insufficient clustering, leaving us with 829 geolocation results.

The highest scoring node in the hierarchy was selected as the geographic focus of a blog. However, if the highest scoring node was a country or a state or province, the highest scoring result that was subsumed by the top scoring result was selected. For example, if Maryland was the top scoring result, the next highest scoring result in the list that was also in Maryland was selected. If there were no other results in Maryland, then Maryland itself was selected. Of the 829 results, 808 were correctly identified as being in the United States, giving us an accuracy of 97% at the country level. We were more interested in accuracy at a lower level of resolution, however, so we applied acceptance criteria to our results that took into consideration the proximity to a blog's

TABLE I. TABLE I Results by Selection Criteria

Acceptance Criteria	Total Hits	Accuracy
Result with 100 miles of known location	409	49%
Result subsumes known location	86	10%
Result in same state/province as known location	31	4%
All criteria	526	63%

known location. The criteria for an acceptable result were defined as being when the extracted geographic focus subsumed the blog's true location, was within 100 miles of it, or was in the same state or province. Using these criteria, we had 526 matches out of 829 usable results, for 63% accuracy. The breakdown of the acceptable results by acceptance criterion is shown in Table 1. The

breakdown of the acceptable results by administrative level is shown in Table 2.

V. Discussion

Our results suggest that for many blogs, the extracted geographic focus does indeed correspond to the author's location. Although our accuracy was comparable to that reported in [1 and 9], our work diverges from their stated goals. They were concerned with extracting the geographic focus of individual Web pages and assigning a geographic location as a topical feature of a page or a subsegment of a page. We were specifically looking at blog posts, taking into the account the personal nature of a large proportion of blog content, aggregating geographic references across a large number of a blog's posts, and then testing whether the extracted geographic focus matched a blogger's location. It is difficult to gauge what an acceptable level of accuracy is for this task. There was a strong agreement, however, between the percentage of disambiguated locations extracted from a blog's posts that met the acceptance criteria described in the last section and whether we had a positive result. The Pearson's correlation coefficient between the paired values (correct result or not, and the percentage of disambiguated places meeting the acceptance criteria) for the 829 blogs was 0.51. This suggests a correlation between the number of mentions of nearby locations and whether or not the blog can be geolocated, and it supports our intuition that if a person frequently mentions locations near where they live, the extracted geographic focus is more likely to coincide with their home location. The accuracy of our prediction, then, is primarily a function of a blogger's style and of how much they reveal in their posts about their location.

Because this method is sensitive to the quality and quantity of disambiguated place names, a number of approaches could be attempted to boost our accuracy. The model that was provided with the Stanford NER was trained on U.S. and U.K. newswire sources from the CoNLL, MUC6, MUC7, and ACE evaluations, so a model trained on blog text may provide a more accurate set of location entity mentions. Applying some of the disambiguation heuristics described in Section II that use document level context might boost disambiguation performance. 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.

TABLE II. Results by Adminstrative Level

Administrative Level	Number of Results	Accuracy
Country (United States)	11	1.3%
State	131	15.8%
County	372	44.9%
City/Populated Place	12	1.4%

Different strategies for the initial scoring of the disambiguated locations were tried, and assigning the same score to all disambiguated nodes gave the best accuracy. Scoring strategies were applied that were based on the confidence associated with the heuristic used to disambiguate a location and on the counts or frequency of a location mention. Neither approach resulted in an accuracy result better than that attained by assigning equal values for the initial scores.

VI. Future Work

There are additional research topics that we plan on pursuing in relation to this work. We did not perform a genre analysis of the blogs in our test set, but an area of future research may be to determine if this technique is sensitive to whether the blog was a personal diary blog, a professional blog, or a blog devoted to a specific topic. There is a rich literature in linguistics and cognitive science on locative language [6], and it might be interesting to build on this and look at techniques that can extract references to proximate locations based on linguistic cues. This would be a way to narrow candidate place names down to those that are likely to refer to nearby locations.

VII. Conclusion

For some subset of blogs, we have demonstrated that the location mentions in the text of posts can be used to determine the author's location. Techniques such as the one we have described that do not use IP or domain based geolocation techniques or geolocating metadata are useful because most blogs are not self hosted and the use of metadata is not widespread.

References

[1] Amitay, E., Har'El, N., Sivan, R., & Soffer, A. (2004). Web-a-where: Geotagging Web content. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 273–280). New York: Association for Computing Machinery.

[2] Finkel, J., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (pp. 363–370). Morristown, NJ: Association for Computational Linguistics.

[3] Kolari P., Java A., Finin, T., Oates T., & Joshi, A. (2006), Detecting spam blogs: A machine learning approach. In *Proceedings of the 21st National Conference on Artificial Intelligence* (pp. 1351-1356). Menlo Park, CA: Association for the Advancement of Artificial Intelligence.

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

[5] Leidner, L. L., Sinclair, G., & Webber, B. (2003). Grounding spatial named entities for information extraction and question answering. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of*

Geographic References (pp. 31–38). Morristown, NJ: Association for Computational Linguistics.

[6] Olivier, P., & Gapp, K.P. (1998). *Representation and processing of spatial expressions*. Mahwah, NJ: Erlbaum.

[7] Overell, S.E., & Ruger, S. (2006). Identifying and grounding descriptions of places. In *SIGIR Workshop on Geographic Information Retrieval* (pp. 14–16). New York: Association for Computing Machinery.

[8] Smith, D., & Crane, G. (2001). Disambiguating geographic names in a historical digital library. In *Proceedings of the Fifth European Conference on Research and Advanced Technology for Digital Libraries* (pp. 127–136). Berlin: Springer

[9] Zong, W., Wu, D., Sun, A., Lim, E. P., & Goh, D. H. (2005). On assigning place names to geography related Web pages. In *Proceedings of the Fifth ACM/IEEE Joint Conference on Digital Libraries* (pp. 354–362). New York: Association for Computing Machinery.

[10] Universal McCann. (2008). *Power to the people—Social Media Tracker Wave 3*. Retrieved on January 13, 2009, from http://www.universalmccann.com/Assets/wave_3_20080403093750.pdf.