Generating Spatial Footprints from Hiking Blogs

Elise Acheson, Flurina M. Wartmann and Ross S. Purves

Abstract Explicitly linking text documents to geographical space is an important processing step for applications such as map visualization, spatial querying, and placename disambiguation. In this work, we present a proof-of-concept processing pipeline to generate spatial footprints for a spatially-rich, manually-annotated corpus of hiking blogs. We present preliminary results obtained by exploiting the spatially-focused nature of our input data and the rich placename resources at our disposal. Future work will fully automate the pipeline and systematically examine the influence of processing decisions on the footprints and downstream tasks.

Keywords Footprints · Toponyms · Document scope · Clustering · Geographic information retrieval

1 Introduction

How can we geographically represent text documents? For documents strongly linked to geographical space, such as hiking blogs, a representative geographical area can be automatically generated by combining natural language processing methods with geographical processing such as spatial filtering or clustering. The resulting document representations, known as 'document scopes' or 'document footprints', are useful in downstream tasks (e.g. spatial queries), upstream tasks (e.g. improving placename disambiguation), and as an end in themselves (e.g. visualizing a text doc-

E. Acheson (✉) · F.M. Wartmann · R.S. Purves
Geography Department, University of Zurich, Winterthurerstrasse 190,
8057 Zurich, Switzerland
e-mail: Elise.Acheson@geo.uzh.ch

F.M. Wartmann
e-mail: Flurina.Wartmann@geo.uzh.ch

R.S. Purves
e-mail: Ross.Purves@geo.uzh.ch

P. Fogliaroni et al. (eds.), *Proceedings of Workshops and Posters at the 13th International Conference on Spatial Information Theory (COSIT 2017)*, Lecture Notes in Geoinformation and Cartography, https://doi.org/10.1007/978-3-319-63946-8_2

ument on a mapping interface) (Monteiro et al. 2016; Purves et al. 2007; Quercini et al. 2010).

The context of our work is a project on how people describe landscapes in Switzerland. With the goal of comparing landscape descriptions from different data sources (hiking blogs and Flickr photos), document footprints are generated for a web-crawled corpus of hiking blogs in order to query and select Flickr photos based on location. For this task, we aimed to generate high-precision, geographically focused footprints.

2 Data and Methods

Our corpus consisted of web documents related to ten study sites in the German speaking region of Switzerland in a first-person narrative. Documents were collected by targeted web-crawling, with five texts per site selected by manual triage, for a final corpus of 50 documents.

To generate document footprints, we followed the established three-step processing pipeline (Amitay et al. 2004; Monteiro et al. 2016) consisting of: (1) identifying placenames, (2) grounding placenames, and (3) generating a footprint (geometry) (Fig. 1). Poor placename identification has been identified as a major source of error for document scope propagated downstream (Amitay et al. 2004; Purves et al. 2007). Thus, to obtain precise footprints, we performed step 1 manually, which was feasible due to the small corpus size. Step 2, grounding, involved querying an API to obtain ranked results from the SwissNames3D gazetteer for each placename, after having aggregated placenames repeating within a study site and recording their frequencies. For the final step, footprint generation, we experimented with two approaches: iterative filtering based on the centroid and standard deviation of our candidate points (Smith and Crane 2001), and clustering using DBSCAN to identify one main cluster and discard outliers.

Finally, we experimented with permutations in processing decisions in order to generate optimal footprints which suited our requirements. Decisions included: how many candidates for each placename to retain at the grounding stage; whether to treat with higher priority placenames with exactly one candidate; and whether to use the

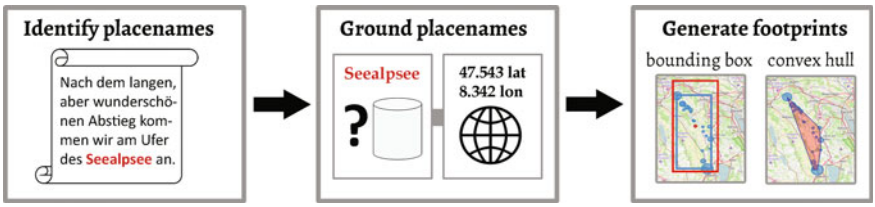

Fig. 1 Processing pipeline for document footprint generation

frequency per site of placenames. We automated the entire processing pipeline, starting from the manually annotated placenames, to output ten convex hulls or bounding boxes on each run.

3 Results and Conclusion

Our preliminary results showed that for placename grounding, simple approaches worked well enough for our purposes: ranked candidate placenames from SwissNames3D were sufficiently accurate that we obtained good results by retaining just the top candidate for each placename. For footprint generation, satisfactory results were obtained using both distance-based filtering and DBSCAN clustering, but the DBSCAN results were better suited to more complex geometric arrangements.

Our footprint requirements stemmed from our downstream tasks: performing a spatial query for Flickr photos, and ultimately, comparing datasources about landscapes at our ten study sites. These task-based requirements, along with the availability of quality placename resources for our area of study, influenced our processing decisions at every stage. Future work will fully automate the placename identification stage, and will systematically measure the effects of permutations in processing decisions on the downstream tasks of querying and document comparison.

References

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, ACM, New York, NY, USA, SIGIR '04, pp 273–280. doi: 10.1145/1008992.1009040

Monteiro BR, Davis CA Jr, Fonseca F (2016) A survey on the geographic scope of textual documents. Comput Geosci. doi:10.1016/j.cageo.2016.07.017. http://www.sciencedirect.com/science/article/pii/S0098300416301972

Purves RS, Clough P, Jones CB, Arampatzis A, Bucher B, Finch D, Fu G, Joho H, Syed AK, Vaid S, Yang B (2007) The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. Int J Geogr Inf Sci 21(7):717–745. doi:10.1080/13658810601169840

Quercini G, Samet H, Sankaranarayanan J, Lieberman MD (2010) Determining the spatial reader scopes of news sources using local Lexicons. In: Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems, ACM, New York, NY, USA, GIS '10, pp 43–52. doi:10.1145/1869790.1869800

Smith DA, Crane G (2001) Disambiguating geographic names in a historical digital library. In: Constantopoulos P, Slvberg IT (eds) Research and advanced technology for digital libraries, no. 2163 in Lecture notes in computer science. Springer, Berlin, Heidelberg, pp 127–136. http://link.springer.com/chapter/10.1007/3-540-44796-2_12