frontiers in Big Data

PERSPECTIVE
published: 12 June 2019
doi: 10.3389/fdata.2019.00012

Identifying Travel Regions Using Location-Based Social Network Check-in Data

*Avradip Sen and Linus W. Dietz**

Department of Informatics, Technical University of Munich, Munich, Germany

Travel regions are not necessarily defined by political or administrative boundaries. For example, in the Schengen region of Europe, tourists can travel freely across borders irrespective of national borders. Identifying transboundary travel regions is an interesting problem which we aim to solve using mobility analysis of Twitter users. Our proposed solution comprises collecting geotagged tweets, combining them into trajectories and, thus, mining thousands of trips undertaken by twitter users. After aggregating these trips into a mobility graph, we apply a community detection algorithm to find coherent regions throughout the world. The discovered regions provide insights into international travel and can reveal both domestic and transnational travel regions.

Keywords: data-mining, human mobility modeling, spatial clustering, region detection, visualization

OPEN ACCESS

Edited by:
Roberto Interdonato, Télédétection et Information Spatiale (TETIS), France

Reviewed by:
Cristian Molinaro, University of Calabria, Italy
Sabrina Gaito, University of Milan, Italy

***Correspondence:**
Linus W. Dietz
linus.dietz@tum.de

Specialty section:
This article was submitted to Data Mining and Management, a section of the journal Frontiers in Big Data

Received: *24 March 2019*
Accepted: *27 May 2019*
Published: *12 June 2019*

Citation:
Sen A and Dietz LW (2019) Identifying Travel Regions Using Location-Based Social Network Check-in Data. Front. Big Data 2:12. doi: 10.3389/fdata.2019.00012

1. INTRODUCTION

The destinations visited within a trip may overarch existing administrative divisions of provinces, federal states, and countries. For example, visiting the Alps of Europe, one is not restricted in travel by country borders as all adjacent countries are members of the Schengen Area. When developing a travel region recommender system for composite trips this is a challenge, because one needs a region model to choose the recommendations from Dietz (2018). To come up with such a model, we propose to observe traveler mobility behavior, aggregate it using spatial clustering methods, thereby re-drawing the boundaries of the world's travel regions using a data-driven approach.

Data collected from location-based social networks has previously been used as a proxy for human mobility, however, such data sets are either not readily available, are focused on small areas, such as cities, or have too sparse check-ins of the users. Hence, we use public Twitter APIs to collect traveler data in the form of geotagged tweets. From the series of tweets, we determine the home location of the user and then extract the trips (Dietz et al., 2018). These trips are then aggregated into a weighted graph of tourist flows with nodes being cities and edges being the number of trips from one city to another. This graph is then fed into a community detection algorithm (Bohlin et al., 2014), whose results constitute the world's travel regions irrespective of established political boundaries.

In this position paper, we want to motivate this approach, describe our ideas to implement and evaluate such a method. Furthermore, we outline the implications and benefits of a data-driven region model in other domains, such as recommender systems.

2. METHOD

Twitter allows algorithmic access to a stream of public tweets through their APIs, which can be queried to build a data set of geotagged tweets. By querying timelines of users who have enabled sharing the geolocation of their tweets, we can follow their movement patterns. To reduce noise, the individual geolocations are matched to the nearest city. Thus, each tweet in the timeline constitutes a check-in to a city. After the home city of the user has been determined by the highest number of check-ins, consecutive check-ins outside of the home city can then be combined to a trip. To focus on travelers, we exclude all trips shorter than 7 days. Furthermore, we require at least one check-in within 5 days, to ensure sufficient data quality. For more details on the trip mining, we refer to our previous paper (Dietz et al., 2018).

The trips are then transformed into an undirected graph, where each city is a node, and the edges represent the flows divided by the distance between the two cities. The flows are computed by summing up the co-occurrences of the two nodes in a clique formed by all cities in a trip. For example, if somebody traveled from Munich to Berlin via Nuremberg in one trip, we would also count the flow from Munich to Berlin as one. Including the distance into the edge weight was useful to reduce noise in the flow graph introduced by distant traffic hubs, such as airports. With this graph-based representation, we can run the Infomap multi-level community detection algorithm to see which cities form coherent clusters (Rosvall et al., 2009).

3. PRELIMINARY RESULTS

Running this approach with trips from Twitter reveals four major clusters on the highest hierarchy:

1. North and Central America,
2. South America,
3. Europe, Russia, Arabia, Western and South Africa, and
4. Eastern Africa, Asia, and Oceania.

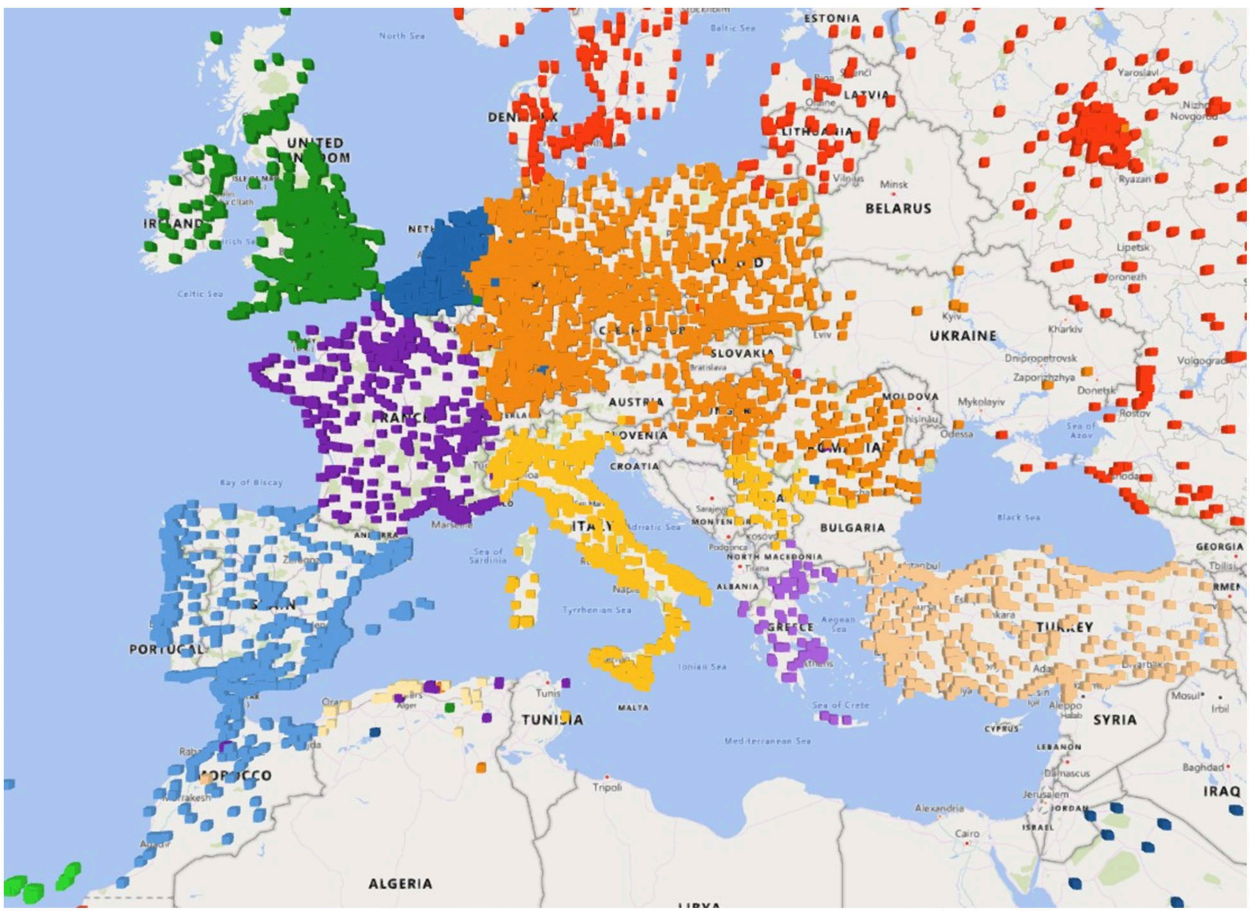

FIGURE 1 | The second-level community structure of Europe.

The level two clusters of Europe, depicted in **Figure 1**, correspond to groups of similar countries. The British Isles, the Iberian Peninsula, and much of Central and Eastern Europe are merged into respective clusters, while countries like France, Italy, and Turkey roughly retain their own clusters. This is already an interesting result, as it shows that political boundaries have a strong influence on the travel behavior. Subdividing these clusters reveals further regions, however the results become more fuzzy and subject to thorough evaluation. One major challenge is to find a termination criterion to decide whether to continue splitting these clusters. In our opinion, this cannot be decided with the current data, but requires further analysis of the regions, such as the number of cities and the area covered. An evaluation of the quality of the discovered region will also prove to be challenging. However, comparing our third-level clusters of the United Kingdom with those of Ratti et al. (2010) revealed high similarities.

4. RELATED WORK

Human mobility analysis has helped us to improve our understanding of traffic forecasting (Kitamura et al., 2000), the spread of diseases (Eubank et al., 2004), and also computer viruses (Kleinberg, 2007). Researchers have already attempted to define regions based on human mobility data for various purposes such as administrative region discovery (del Prado and Alatrista-Salas, 2016), topical region discovery (Taniguchi et al., 2015), and political redistricting (Joshi et al., 2009). Closest to our approach is the work of Hawelka et al. (2014), who aim to find larger regions of mobility, by combining several countries. We aim to find touristic regions that are smaller and potentially independent of countries.

There are various algorithms to perform spatial clustering and community detection, such as the Louvain method (Blondel et al., 2008), GDBSCAN (Ester et al., 1996), and Infomap (Rosvall et al., 2009). They are comparable in runtime complexity, however (Fortunato and Hric, 2016) finds that the Infomap algorithm outperforms the Louvain method in the quality of the communities. GDBSCAN uses the distance between points explicitly to form clusters that are geographically contiguous. Thus, we use Infomap, as it allows to use self-computed weights for the graph and can detect hierarchies. This resolves the resolution limit problem, where the size of communities depend on the size of the graph, which can result in recognized communities being merged together in large networks.

5. CONCLUSIONS

This position paper introduces an approach for spatial clustering of touristic regions from trips mined from Twitter. To the best of our knowledge, this is the first application of geo-located tweets to find travel regions, with data spanning the whole world. The analysis of results finds a coherent hierarchy of clusters. This confirms that the use of tweets to find traveler mobility patterns and define regions based on the patterns is a feasible approach.

In future, we plan to make a thorough evaluation of the resulting regions using numeric method, but also to visually compare them to findings of other region discovery approaches.

DATA AVAILABILITY

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

AS: Prototype implementation, experimentation, literature analysis. LD: Main author of manuscript, developed the trip mining library.

FUNDING

This work was supported by the German Research Foundation (DFG) and the Technical University of Munich (TUM) in the framework of the Open Access Publishing Program.

REFERENCES

Blondel, V. D., Guillaume, J. L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech. Theor. Exp.* 2008, 1–12. doi: 10.1088/1742-5468/2008/10/P10008

Bohlin, L., Edler, D., Lancichinetti, A., and Rosvall, M. (2014). "Community detection and visualization of networks with the map equation framework," in *Measuring Scholarly Impact: Methods and Practice*, eds Y. Ding, R. Rousseau, and D. Wolfram (Cham: Springer), 3–34.

del Prado, M. N., and Alatrista-Salas, H. (2016). "Administrative regions discovery based on human mobility patterns and spatio-temporal clustering," in *Proceedings of 13th International Conferences on Mobile Ad Hoc and Sensor Systems, MASS'16* (Brasilia: IEEE), 65–74.

Dietz, L. W. (2018). "Data-driven destination recommender systems," in *Proceedings of 26th Conferences User Modeling, Adaptation and Personalization, UMAP '18* (New York, NY: ACM).

Dietz, L. W., Herzog, D., and Wörndl, W. (2018). "Deriving tourist mobility patterns from check-in data," in *Proceedings of the WSDM WS on Learning from User Interactions* (Los Angeles, CA).

Ester, M., Kriegel, H. P., Sander, J., and Xu, X. (1996). "Density-Based Clustering Methods," in *KDD-96 Proceedings* (Portland, OR).

Eubank, S., Guclu, H., Kumar, V. S. A., Marathe, M. V., Srinivasan, A., Toroczkai, Z., et al. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature* 429, 180–184. doi: 10.1038/nature02541

Fortunato, S., and Hric, D. (2016). Community detection in networks. *Phys. Rep.* 659, 1–44. doi: 10.1016/j.physrep.2016.09.002

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., and Ratti, C. (2014). Geo-located twitter as proxy for global mobility patterns. *Cartogr. Geogr. Inform. Sci.* 41, 260–271. doi: 10.1080/15230406.2014.890072

Joshi, D., Soh, L.-K., and Samal, A. (2009). "Redistricting using heuristic-based polygonal clustering," in *Proceedings of 9th International Conference on Data Mining* (Miami, FL: IEEE), 830–835.

Kitamura, R., Chen, C., Pendyala, R. M., and Narayanan, R. (2000). Micro-simulation of daily activity-travel patterns for travel demand forecasting. *Transportation* 27, 25–51. doi: 10.1023/A:1005259 324588

Kleinberg, J. (2007). Computing: the wireless epidemic. *Nature* 449, 287–288. doi: 10.1038/449287a

Ratti, C., Sobolevsky, S., Calabrese, F., Andris, C., Reades, J., Martino, M., et al. (2010). Redrawing the map of Great Britain from a network of human interactions. *PLoS ONE* 5:e14248. doi: 10.1371/journal.pone.0014248

Rosvall, M., Axelsson, D., and Bergstrom, C. T. (2009). The map equation. *Eur. Phys. J. Spec. Top.* 178, 13–23. doi: 10.1140/epjst/e2010-01179-1

Taniguchi, Y., Monzen, D., Ariestien, L. S., and Ikeda, D. (2015). "Discover overlapping topical regions by geo-semantic clustering of tweets," in *29th International Conference Advanced Information Networking and Applications Workshops* (Gwangiu: IEEE), 552–557.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor and reviewer SG declared their involvement as co-editors in the Research Topic, and confirm the absence of any other collaboration.