Text Meets Space: Geographic Content Extraction, Resolution and Information Retrieval

Jochen L. Leidner[1,2,3](✉), Bruno Martins[4], Katherine McDonough[5,6], and Ross S. Purves[7]

[1] Polygon Analytics Ltd., 19a Canning Street, Edinburgh EH3 8HE, Scotland, UK
[2] Refinitiv Labs, Refinitiv Ltd., 5 Canada Square, London E14 5AQ, UK
[3] Regents Court, University of Sheffield, 211 Portobello, Sheffield S1, UK
leidner@acm.org
[4] Computer Science and Engineering Department, IST, University of Lisbon, Lisbon, Portugal
[5] British Library, Alan Turing Institute, 96 Euston Rd, London NW1 2DB, UK
[6] School of History, Queen Mary University, London, UK
[7] Department of Geography, University of Zurich, Zurich, Switzerland

Abstract. In this half-day tutorial, we will review the basic concepts of, methods for, and applications of geographic information retrieval, also showing some possible applications in fields such as the digital humanities. The tutorial is organized in four parts. First we introduce some basic ideas about geography, and demonstrate why text is a powerful way of exploring relevant questions. We then introduce a basic end-to-end pipeline discussing geographic information in documents, spatial and multi-dimensional indexing [19], and spatial retrieval and spatial filtering. After showing a range of possible applications, we conclude with suggestions for future work in the area.

1 Introduction

The notion of geographic relevance and the role of geographic space in information access have been recognized for a long time [15]. For example, the PERSEUS digital library aimed to make humanities documents accessible spatially, while e.g. the SEQUOIA and SPIRIT projects [8], as well as the GeoCLEF shared task [1] aimed to study geographic information retrieval. More recently, the pervasiveness of mobile computing devices [11] and other developments associated to the Internet of Things (IoT) all necessitate reflection on the role of geographic space in making information collected and stored accessible, not just indexed using words and numbers but also spatially. However, to date, not ECIR nor other IR conferences have offered a tutorial for interested researchers and practitioners, making the body of research that make up the state of the art accessible.

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12036, pp. 669–673, 2020.
https://doi.org/10.1007/978-3-030-45442-5_89

To this end, we propose a half-day to address this gap. We will introduce or recap the core concepts from geography and its intersection with IR, and survey existing techniques to (a) construct spatial representations from textual documents and queries (typically exploiting geographic knowledge from gazetteers [7] in doing so), and to (b) utilize geographic knowledge (prior and extracted from data) to better access document collections in which geographic space place a substantial roles. We will also cover example applications [5], e.g. in fields such as the digital humanities [12], and discuss possible avenues for future work in the area.

2 Goals and Objectives

In this tutorial, we aim to give a survey of the concepts and methods used to make implicit spatial evidence contained in text collections accessible. We cover selected early and seminal attempts [3,8,10,13] and more recent Machine Learning (ML) methods [6,16–18], hoping to inspire students and fellow researchers to get interested in conducting their own research in this area. Bringing two seemingly disparate worlds like geographic space and text documents together is exciting!

By the end of the tutorial session, the attendees will have a clear sense of the key concepts in Geo-NLP and Geographical Information Retrieval (GIR), and they will understand some seminal methods as well as open problems.

3 Description and Structure of the Tutorial

This one day tutorial will be divided into five sessions:

– Geography and text: an introduction to the ways in which geographic concepts are reflected in natural language and in text;
– Toponym recognition and resolution [9]: key to most geographically inspired analysis are the use of place-names in text, their identification, disambiguation, and resolution to unique locations;
– Geographic relevance and ranking [4]: methods for incorporating geographic information in IR indexes and ranking algorithms. Discuss what is geographic relevance, and how it varies with context and application domain;
– Applications: Concrete examples for the application of the introduced methods, in fields ranging from Digital Humanities to Web search, together with a discussion on requirements and their implications on algorithmic and data choices;
– Future challenges: Where are the most likely applications of GIR in the future, and what are key societal and methodologically driven challenges;

The first four sessions will each present fundamental challenges, a selection of examples from the state of the art, and include interactive exercises (computer and/or paper based) to illustrate basic concepts to participants.

4 Prerequisites

In terms of prerequisites, some knowledge of basic IR and ML concepts will be helpful. However, the tutorial is designed for a broad audience, introducing key high level concepts, and providing participants with material to deepen knowledge subsequently.

5 Target Audience

The target audience for this tutorial includes the following three groups:

- students of computer science, especially in information retrieval, who want to learn about mobility-relevant spatial computation around search/IR (e.g. [2]);
- practicing IR engineers who would like to expand their areas of expertise so as to include geographic search;
- information retrieval researchers interested in and introduction and state-of-the-art review [14] on GIR and Geo-NLP;
- geographers or GIS experts who have not yet worked with text, and who would like to learn how the spatial knowledge implicit in text collections can be used to support geospatial analysis.

Beyond these directly targeted groups, the tutorial could be of interest to anyone who would like to understand better how the world of geographic space relates to the world of unstructured textual documents.

6 Presenters and Their Experience

Jochen L. Leidner is a computer scientist and research manager. He is Director of Research at Refinitiv Labs (formerly Thomson Reuters F&R) in London where he leads the Research & Development function and team. A computational linguist by training, he holds Master's degrees (Erlangen and Cambridge) and a Ph.D. (Edinburgh). His 2007 Ph.D. thesis "Toponym Resolution in Text" (published in book form in 2008) attracted over 200 hundred citations. He is a Fellow of the Royal Geographical Society and currently also the Academy of Engineering Visiting Professor of Data Analytics in the Department of Computer Science at the University of Sheffield to instill industry practice into engineering training.

Bruno Martins is an assistant professor at the Department of Computer Science and Engineering of Instituto Superior Técnico in the University of Lisbon and a researcher at INESC-ID, where he works on problems related to the general areas of information retrieval, text mining, and the geographical information sciences. He has been involved in several research projects related to geospatial aspects in information access and retrieval, and he has accumulated a significant expertise in addressing challenges at the intersection of information retrieval, machine learning, and the geographical information sciences.

Katherine McDonough is a Senior Research Associate at The Alan Turing Institute with the Living with Machines project and a Research Fellow at Queen Mary, University of London. She has formerly taught and worked on digital humanities projects at Stanford University, Western Sydney University, and Bates College. With a background in eighteenth-century French history, her early research focused on the politics of infrastructure. She has written on GIR challenges for humanities research and is a member of the GéoDisco project, which examines geographic discourse in historical French encyclopedias. Her current work explores new approaches to GIR informed by humanistic source criticism.

Ross Purves is a professor at the University of Zurich. His research focuses on the geographic analysis of text, exploring both methodological issues (e.g. gazetteer quality and representation of vernacular names) and analysis of text to better understand landscape. He collaborated on the SPIRIT project, which investigated a number of concepts fundamental to geographic information retrieval. Together with Chris Jones, he organises the workshop on Geographic Information Retrieval which has been hosted by CIKM, SIGIR and ACM SIGSPATIAL, and which has been an important incubator of many ideas related to GIR. He recently co-authored a comprehensive review of GIR [14].

7 Previous Events

This is a new tutorial, and therefore was never presented before. All of the presenters are experienced teachers and have given seminars at a range of international conferences on related material.

8 Summary and Conclusion

We have presented a tutorial proposal for geospatial content processing and retrieval. Geographic aspects in information access and retrieval have been increasing in relevance, given the interest in analysing huge volumes of unstructured data in fields such as the digital humanities or the computational social sciences, and given the pervasiveness of networked sensors, GPS-enabled mobile devices, and in-car navigation systems. Modern information systems need to spatially enable text to make it accessible to a variety of use cases that contain a notion of "geographic relevance". This suggests that our novel tutorial would be likely to be of interest to most attendees of ECIR 2020.

References

1. GeoCLEF (n.d.). http://www.clef-initiative.eu/track/geoclef. Accessed 2019
2. Al-Olimat, H.S., Shalin, V.L., Thirunarayan, K., Sain, J.P.: Towards geocoding spatial expressions. Technical report (unpublished). https://arxiv.org/pdf/1906.04960.pdf. Accessed 2019

3. Amitay, E., Har'El, N., Sivan, R., Soffer, A.: Web-a-where: geotagging web content. In: SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, 25–29 July 2004, pp. 273–280 (2004)
4. Andogah, G.: Geographically Constrained Information Retrieval. Ph.D. thesis, University of Groningen, Groningen, The Netherlands (2010)
5. Ding, J., Gravano, L., Shivakumar, N.: Computing geographical scopes of web resources. In: El Abbadi, A., et al. (eds.) VLDB 2000, Proceedings of 26th International Conference on Very Large Data Bases, 10–14 September 2000, Cairo, Egypt, pp. 545–556. Morgan Kaufmann (2000)
6. Gritta, M., Pilehvar, M.T., Collier, N.: Which Melbourne? Augmenting geocoding with maps. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 1285–1296. Association for Computational Linguistics, Melbourne (2018)
7. Hill, L.: Georeferencing. MIT Press, Cambridge (2009)
8. Joho, H., Sanderson, M.: The SPIRIT collection: an overview of a large web collection (2004)
9. Leidner, J.L.: Toponym Resolution in Text. Universal Press, Irvine (2008)
10. Leidner, J.L., Sinclair, G., Webber, B.: Grounding spatial named entities for information extraction and question answering. In: Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References held at HLT/NAACL 2003, pp. 31–38 (2003). https://www.aclweb.org/anthology/W03-0105
11. Mathew, W., Raposo, R., Martins, B.: Predicting future locations with hidden markov models. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, UbiComp 2012, pp. 911–918. ACM, New York (2012)
12. McDonough, K., Moncla, L., van de Camp, M.: Named entity recognition goes to old regime France: geographic text analysis for early modern French corpora. Int. J. Geogr. Inf. Sci. **33**(12), 2498–2522 (2019)
13. Overell, S.E., Rüger, S.M.: Using co-occurrence models for placename disambiguation. Int. J. Geogr. Inf. Sci. **22**(3), 265–287 (2008)
14. Purves, R.S., Clough, P., Jones, C.B., Hall, M.H., Murdock, V.: Geographic information retrieval: progress and challenges in spatial search of text. Found. Trends Inf. Retrieval **12**(2–3), 164–318 (2018)
15. Sanderson, M., Kohler, J.: Analyzing geographic queries. In: Proceedings of the Workshop on Geographical Information Retrieval held at SIGIR 2004. http://www.geounizh.ch/~rsp/gir/
16. Santos, R., Murrieta-Flores, P., Calado, P., Martins, B.: Toponym matching through deep neural networks. Int. J. Geogr. Inf. Sci. **32**(2), 324–348 (2018)
17. Speriosu, M., Baldridge, J.: Text-driven toponym resolution using indirect supervision. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 1466–1476. Association for Computational Linguistics, Sofia (2013)
18. Yan, B., Janowicz, K., Mai, G., Gao, S.: From ITDL to Place2Vec: reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. In: Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL 2017, pp. 35:1–35:10. ACM, New York (2017)
19. Zhang, X., Du, Z.: Spatial indexing. In: Wilson, J.P. (ed.) The Geographic Information Science & Technology Body of Knowledge. UCGIS, 4th Quarter 2017 Edition (2017). https://gistbok.ucgis.org/bok-topics/spatial-indexing