The Lifecycle of Geotagged Data

Rossano Schifanella
University of Turin
Turin, Italy
schifane@di.unito.it

Bart Thomee
Google
San Bruno, CA, USA
bthomee@google.com

David A. Shamma
Centrum Wiskunde & Informatica
Amsterdam, Netherlands
aymans@acm.org

ABSTRACT

The world is a big place. At any given instant something is happening somewhere, but even when nothing in particular is going on people still find ways to generate data, such as posting on social media, taking photos, and issuing search queries. A substantial number of these actions is associated with a location, and in an increasingly mobile and connected world (both in terms of people and devices), this number is only bound to get larger. Yet, in the literature we observe that many researchers often unwittingly treat the geospatial dimension as if it were a regular feature dimension, despite it requiring special attention. In order to avoid pitfalls and to steer clear of erroneous conclusions, our tutorial aims to teach researchers and students how geotagged data differs from regular data, and to educate them on best practices when dealing with such data. We will cover the lifecycle of how geotagged data is used in research, where the topics range from how it is *created*, *represented*, *processed*, *modeled*, *analyzed*, *visualized*, and *perceived*. The tutorial requires both passive and active involvement—we not only present the material, but the attendees also get the opportunity to interact with it using a variety of open source data and tools that we have prepared using a virtual machine. Attendees should expect to leave the course with a high-level understanding of methods for properly using geospatial data and reporting results, the necessary context to better understand the geography literature, and resources for further engaging with georeferenced data.

Keywords

Geospatial data; Geotagged data; Spatiotemporality; Tutorial

1. INTRODUCTION

Geography plays an important role in everyday life, and many of the decisions people take depend on where they live, where they are now, and the locations they are familiar with. People's actions are frequently analyzed in the context of research, but the geographic component of their personal circumstances and actions does not always receive the attention it deserves or needs. For instance, while recommending good restaurants obviously depends on someone's food and drink preferences, and how far the restaurants are located from where the person is now, it is easy to forget that natural and man-made barriers can affect the optimal recommendations, see Figure 1.

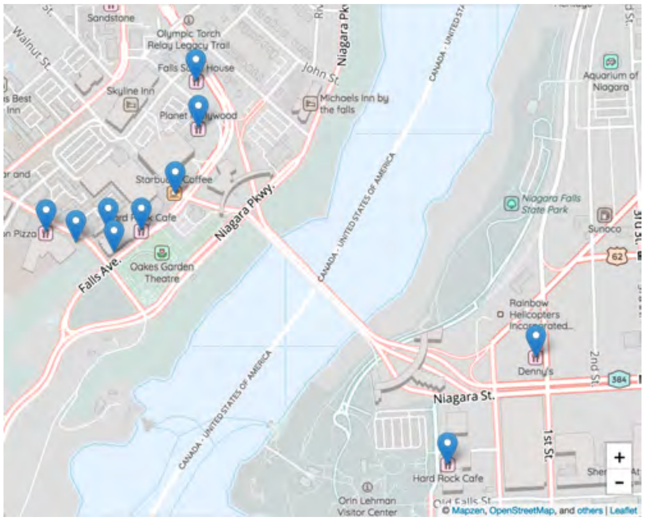

Figure 1: Selection of recommended restaurants around Niagara Falls. It may not be trivial to reach a restaurant located on the other side of the bridge, as that requires crossing the US-Canada border.

The main objective of our tutorial is to arm attendees with both theoretical and practical knowledge about the whole process of making sense of geospatial data, rather than focusing on specific technologies, tools or data sources. Our tutorial aims to provide a broad vision of all the processes and technologies available to researchers, and it centers on the basics of geospatial understanding to convey best practices when dealing with such data. In addition, we present how geospatial data is different from other kinds of data, and therefore requires special considerations when representing, processing, modeling, analyzing and visualizing. We also show how cartography can be used to drive a point home, see Figure 2.

In many instances, treating geospatial data as if it were just standard two-dimensional data works just fine. However, sooner or later the properties of the Earth have to be

WWW 2017, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3051102

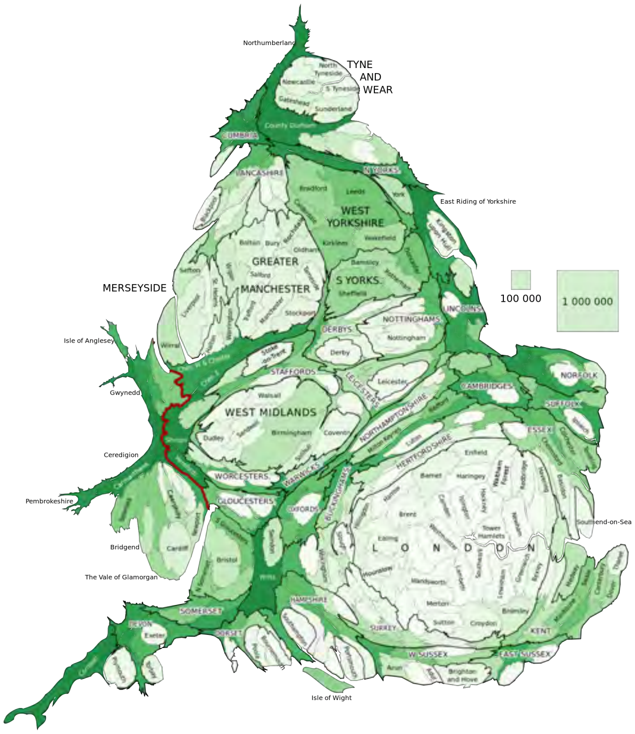

Figure 2: Map of England and Wales in which each administrative area is drawn with a size proportional to their population according to 2011 Census data. The darker the color, the bigger the real area is. Image by PawełS 🅭🅯⊜ `https://goo.gl/aofI6I`.

taken into account to prevent errors from negatively affecting any calculations. For example, while a degree of latitude measures about 111km, a degree of longitude varies in size depending on how close to the equator or the poles it is; using the Euclidean distance to find the nearest neighbor of a point may therefore not yield the correct neighbor, because it treats each degree of latitude the same as a degree of longitude. The dateline also poses interesting problems due to the degrees of longitude switching from −180° to +180°; for instance, when this coordinate flipping is not taken into account, the shortest distance between the eastern and western parts of Fiji traverses the entire globe! With some care it is often possible to build in proper support for geospatial data in algorithms.

2. OUTLINE

The full-day tutorial is organized in 3 modules that cover the main phases of the lifecycle of geotagged data in research, containing both theory and experimental results. Each module is about 2 hours long, where each is divided into approximately 90 minutes of frontal lecture and 30 minutes of interactive session, during which the attendees will play with geotagged data and the tools introduced in the module to solve a real problem.

Module 1: Perception We present an introduction into modern geography theory, where we dive into specific areas of the geography literature that are particularly relevant to the WWW audience. We further describe how machines commonly represent geographic data and how humans in turn perceive this data. This includes discussions on how people discuss and create places beyond simple coordinate reference systems.

Module 2: Analytics We first cover basic techniques for operating on geotagged data, such as determining the distance between geographic coordinates, and computing areas of and overlaps between polygons. We then move to more advanced techniques, such as clustering and density estimation, in order to prepare the data for further analysis. We particularly show how geotagged data differs from traditional data and thus often requires special considerations in order to obtain reliable output, such as understanding which statistical techniques are (not) appropriate for handling geographic data. An important focus is placed on how the data representation influences which techniques should (not) be used. We will let the attendees experience all these facets of processing, modeling, and analyzing geotagged data themselves.

Module 3: Visualize We cover a variety of techniques the attendees can use to visualize and explore actionable insights from geotagged data. A hands-on session will let the attendees first interact with real geotagged data to get familiar with visualizing a number of data representations using projections, and then present several use cases for them to investigate using suitable techniques.

Throughout the modules we highlight tools that can assist the attendees to better understand the data. In particular, since the world is not flat, it is not straightforward to correctly visualize geographic data. In each module we will therefore liberally use visualization techniques to illustrate how geotagged data should be displayed and how this can help understanding. The hands-on sessions will teach the attendees how to effectively use the right tool at the right time to maximize the knowledge they can extract from the data.

3. AUDIENCE

This introductory tutorial targets all researchers and students that want to learn more about how to properly work with geotagged multimedia data. It provides information to get complete novices started, while at the same time does not shy away from presenting advanced representation, modeling and analysis techniques for those interested in a deeper understanding of geographic data. A substantial portion of the data on the World Wide Web refers to specific geographic places or areas, and in an increasingly mobile world this data is created and consumed at varying locations. Considering that hundreds of papers that use geotagged data are published every year, each year more than the year before, we deem our tutorial to be particularly relevant to the audience at the conference.

4. MATERIAL

The tutorial material was published on our tutorial website[1] one week before the event to give time to the attendees to explore the material needed to successfully complete the interactive sessions assignments in advance. The tutorial web

[1] `https://sites.google.com/view/geocycle-www17/`

site contains a general description of the topics covered and for each module it has made the following available:

- Slides presented by the instructor during the lecture.
- Links to external material referred to in the slides.
- Development environment used in the interactive session.

The tutorial web site provides a detailed how-to for setting up the development environment. To minimize the attendees' effort and to have a homogeneous platform for each of them we have made a virtual machine available that provides all the tools, libraries, code, examples, exercises, and data in an easy-to-install cross-platform package. For last minute registrants we had the virtual machine stored on a portable hard disk for them to copy onto their own laptops. All of the material we used is either open source with suitable licenses (e.g. Creative Commons) or in the public domain. Our tutorial does not require internet connectivity, but can benefit from it when available.

5. INSTRUCTORS

Rossano Schifanella is an Assistant Professor in Computer Science at the University of Turin, Italy, where he is a member of the Applied Research on Computational Complex Systems group. He is a former visiting scientist at Yahoo Labs and at the Center for Complex Networks and Systems Research at the Indiana University where he was applying computational methods to model social behavior in online platforms. His research embraces the creative energy of a range of disciplines across technology, computational social science, data visualization, and urban informatics. He is passionate about building new mapping tools that capture the sensorial layers of a city [1, 2], and designing computational frameworks to model aesthetics [4], creativity [3], and figurative language in multimedia platforms.

Bart Thomee is a Software Engineer at Google/YouTube in San Bruno, CA, USA and was previously a Senior Research Scientist at Yahoo Labs and Flickr, where his research focused on the visual and spatiotemporal dimensions of media, in order to better understand how people experience and explore the world, and how to better assist them with doing so [5, 6, 7]. He led the development of the YFCC100M [8] dataset released in 2014, and previously was part of the efforts leading to the creation of both MIRFLICKR datasets. He has furthermore been part of the organization of the ImageCLEF photo annotation tasks 2012–2013, the MediaEval placing tasks 2013–2016, and the ACM MM Yahoo-Flickr Grand Challenges 2015–2016. In addition, he has served on the program committees of, amongst others, ACM MM, ICMR, SIGIR, ICWSM and ECIR. He was part of the Steering Committee of the Multimedia COMMONS 2015 workshop at ACM MM and co-chaired the workshop in 2016; he also co-organized the TAIA workshop at SIGIR 2015.

David A. Shamma is a Principal Research Scientist at Centrum Wiskunde & Informatica. Previously, he was the founding Director of HCI Research at Yahoo Labs and Flickr. His work in social computing and multimedia has been peer review published in over 70 publications and he holds over 12 US and International related patents. His research has been featured in the New York Times, Wired, PetaPixel and Engadget to name a few. He is a member of the ACM MM Steering Committee, the ACM TVX Steering Committee and a Distinguished Member of the ACM. Before Yahoo, he received his Ph.D. from Northwestern University in 2005 and was previously a visiting scientist at NASA's Center for Mars Exploration.

6. REFERENCES

[1] D. Quercia, R. Schifanella, and L. Aiello. The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. In *Proceedings of the ACM Conference on Hypertext and Social Media*, pages 116–125, 2014.

[2] D. Quercia, R. Schifanella, L. Aiello, and K. McLean. Smelly maps: The digital life of urban smellscapes. In *Proceedings of the AAAI International Conference on Weblogs and Social Media*, 2015.

[3] M. Redi, N. O'Hare, R. Schifanella, M. Trevisiol, and A. Jaimes. 6 seconds of sound and vision: Creativity in micro-videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4272–4279, 2014.

[4] R. Schifanella, M. Redi, and L. Aiello. An image is worth more than a thousand favorites: Surfacing the hidden beauty of flickr pictures. In *Proceedings of the AAAI International Conference on Web and Social Media*, 2015.

[5] B. Thomee, I. Arapakis, and D. Shamma. Finding social points of interest from georeferenced and oriented online photographs. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 12(2):36, 2016.

[6] B. Thomee and G. De Francisci Morales. Automatic discovery of global and local equivalence relationships in labeled geo-spatial data. In *Proceedings of the ACM International Conference on Hypertext and Social Media*, pages 158–168, 2014.

[7] B. Thomee and A. Rae. Uncovering locally characterizing regions within geotagged data. In *Proceedings of the IW3C2 International Conference on World Wide Web*, pages 1285–1296, 2013.

[8] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, Jan. 2016.