Building and Mining Large Datasets from Streaming APIs using a Geographic Lens

Michael Martin[1] and Nadine Schuurman[2]

1 Geography, Simon Fraser University, memartin@sfu.ca

2 Geography, Simon Fraser University, nadine@sfu.ca

Research Context

The geoweb is a constantly evolving and mutating place where geographic information is created, analyzed and visualized (Crampton 2008). It has changed from its beginnings in the early 1990s to become the sophisticated web applications of today. Through this evolution, new ways of publishing data online has become easier and easier and now include the wisdom of lay-citizens (Goodchild 2007). The geographic data that citizens provide has taken many shapes, including traditional geographic features found on Open Street Map (Haklay and Weber 2008) and the geo-tagged information of social media (Crooks, Croitoru et al. 2012). The large volumes of information that is being generated through these applications has been termed big data (Elwood, Goodchild et al. 2013) and this research talk will focus on the ways that it can be harvested, analyzed and visualized.

Goals

The goal of this research project is to analyze big data sources using a geographical lens. We plan to establish a geographically enabled database that can acquire data from multiple web sources such as Twitter (www.twitter.com), Flickr (www.flickr.com), and Instagram (www.instagram.com) that can be queried and displayed geographically. Beyond acquiring data, we plan to analyze the data in our database. An example of the type of analysis we would like to complete comes from Twitter (www.twitter.com). Much of Twitter analysis is completed from a lexical standpoint, using computer science methods such as natural language processing to determine sentiment (Go, Bhayani et al. 2009; Pak and Paroubek 2010). This research will build on sentiment methods to specifically look at not only affect in twitter data, but also how geography correlates with sentiment. It is our hope that we will be able to apply the methods we develop from one data source to another.

Methods

At this early stage, our methods are still in development. Currently, these involve writing programs that ingest streaming data via website specific application programming interfaces (APIs) to a MySQL database with spatial extensions. To visualize the data that we have acquired, we have used popular geospatial programs such as QGIS (www.qgis.org) and ArcGIS (www.esri.com).

Anticipated Results

We anticipate that we will be able to create geospatially visualized results of our data analysis algorithms using geospatial web (geoweb) technologies. It is our hope that these visualized results will be a step towards analyzing big data sources with a specifically geographic lens.

Expected Impact

The impact of this work could be relevant in several research fields. We can envisage our algorithms being relevant to social science research as a tool to query specific research topics and social phenomena and see how these vary over space. For industry, we can see how the results of this project may increase the abilities for private companies to understand how their products, or product domains, vary over space and gauge consumer sentiment.

References

Crampton, J. (2008). "Cartography: maps 2.0." Progress in Human Geography.

Crooks, A., A. Croitoru, et al. (2012). "# Earthquake: Twitter as a distributed sensor system." Transactions in GIS.

Elwood, S., M. F. Goodchild, et al. (2013). Prospects for VGI research and the emerging fourth paradigm. Crowdsourcing Geographic Knowledge, Springer**:** 361-375.

Go, A., R. Bhayani, et al. (2009). "Twitter sentiment classification using distant supervision." CS224N Project Report, Stanford: 1-12.

Goodchild, M. (2007). "Citizens as sensors: the world of volunteered geography." GeoJournal **69**(4): 211-221.

Haklay, M. and P. Weber (2008). "OpenStreetMap: User-Generated Street Maps." Pervasive Computing, IEEE **7**(4): 12-18.

Pak, A. and P. Paroubek (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. LREC.