Scalable Twitter Analytics

A Hadoop framework for spatiotemporal social media analytics

Ben Liebersohn

Earlham COLLEGE

CyberGIS

XSEDE
Extreme Science and Engineering Discovery Environment

Abstract

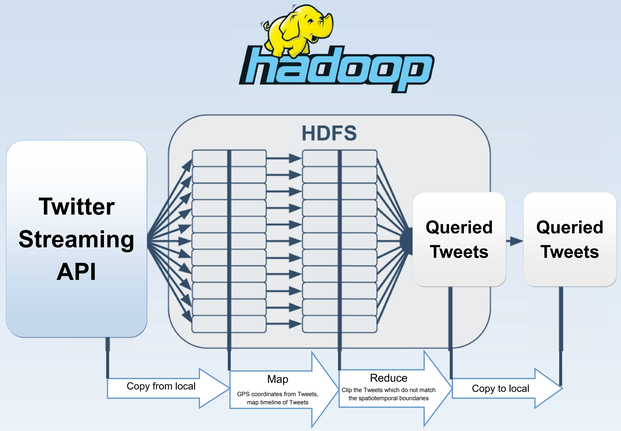
Many researchers are interested in finding better ways of exploring boundaries between physical spaces and digital spaces. Social media users continue the trend towards sharing more information regarding their activities, views, and interests, and even share information relevant to crime monitoring, civil unrest, and human mobility patterns. **With a scalable social media event detection model, we can better extract the relevant interaction between cyber-events and real-world phenomena.** Because over 500 Million Tweets are published annually, sifting through such an enormous database has proved non-trivial. **Through database size reduction, this Hadoop tool makes it possible to examine a reduced size database, making more computationally intensive datamining possible.** First, I define what constitutes a cyber-event. While we could consider all Tweets as cyber-events, for our purposes we will classify Tweets by time and spatial density. By stud-ying daily sets of messages, which combined total over 200 million messages per day, a study of the metadata from certain times and places can be further examined as part of an event, allowing us to run further analysis with a specified event in mind.

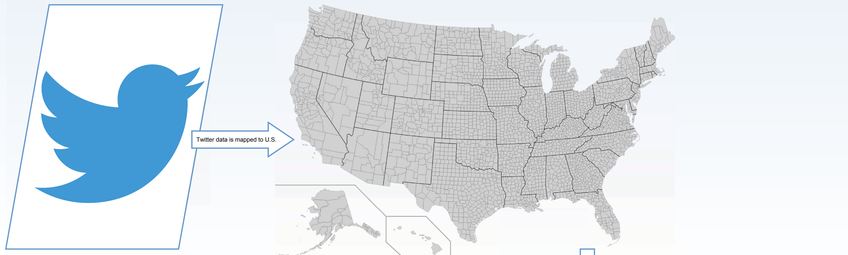
Introduction

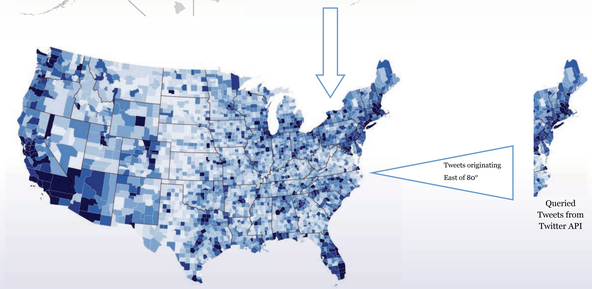
- Spaces exist in many ways topographically. Topographic data extracted from the Twitter streaming API, such as locations and bounding coordinates, can tell us about the behavior of the users who submit the content.
- The data collected using the Twitter API is collected by the CyberGIS Center at the National Center for Supercomputing Applications on the ROGER cyberGIS supercomputer, an XSEDE resource.
- This has been compiling a catalogue of data from a span of 3 years, I can explore current phenomena with a degree of historical context. This dataset, for the purposes of my study, will be looking a Twitter users in the United States, from which the Twitter Streaming API provides approx. 1% of the total throughput traffic in the continental United States. These Tweets provide the basis for an analysis of cyber-events.

Preliminaries

- Twitter gives a bounding box of four sets of GPS coordinates, whose bounding box (w, h) encapsulates the specified place. Larger bounding boxes are less useful.
- My Hadoop code, written in Java, begins by taking user specified time boundaries. In most of my experiments, I selected data by month. The existing HDFS structure I am using makes this task trivial.
- The spatial data is then mapped. Twitter gives a bounding box of four sets of GPS coordinates, whose bounding box (w, h) encapsulates the specified place.
- The Twitter data is first collected on the ROGER cyberGIS supercomputer (an using the Twitter API, and is directly saved into an HDFS style database for examination using Hadoop. The Hadoop framework provides a scalable plat-form for map-reduce style operations.
- Hadoop is used to make a curated database. I am using a reduced dataset which can be parsed with greater stability. By constructing a database with Hadoop, I can ultimately work in a relational database such as PostgreSQL, and are not worry about the size of a database.







Reduced distribution of Twitter users based on Twitter API sample in 2016

Overview

- Once logged into a Hadoop node on ROGER, we instruct Hadoop on how to run the job.
  - supply the location of the data in HDFS (ADDRESS="$HOME/twitter_analysis")
  - set the upper bound of reducers (MAXREDUCERS=64)
  - set the time range DATE by month (DATE=201612 is December, 2016)
- Next, we can begin extracting Tweets. First, let's go over some pseudocode
  - For each file *i* containing Tweets
    - Input IN=$DATE[$*i*], and OUT=tweets/$DATE[$*i*]
    - create generic filesystem *fs* and recursively delete directory
- Now we can run a mapreduce on the code
  - Instruct Hadoop where our bundled code is (hadoop jar $ADDRESS/process.jar), $IN, $OUT, $MAXREDUCERS, bounding GPS coordinates, "true" for messages in our domain
- Last, we recursively delete the *fs* $IN (hadoop fs -rmr $IN)

Results

- In a trial conducted on the ROGER cyberGIS supercomputer, the reduce stage (wherin all locales were mapped) reduced the filesize of the database (beginning on 1/1/2013) from 2.2TB to 1.6GB for just 12/11/2016.
- In a similar trial, Tweets mapped in Indiana and reduced to 12/11/2016 amounted to 28MB.

Conclusions

- Making thematic maps using Twitter data to create new topographic boundaries may yeild new insights into how the users of social media reflect the views of a given region.
- Filesize is approximately days*density/averageUSA_tweetdensity*1.5GB assuming 200,000,000 messages per day.
- Topographic data extracted from my datasets, can tell us about the behavior of the users who submit Tweets in a specified spatiotemporal zone.
- This method of Tweet extraction is a valuable part of cyberGIS data analytics toolkit.
- By filtering my dataset on spatiotemporal parameters, s by consolidating reports with Twitter. In addition, bottlenecks on dense urban communication networks is eased in times of crisis if social media can be a plat-form for detecting disaster.
- Timescales of events not only unify information collection in a world of geospatial differences
- By focusing on certain categories of stories, we can generate metadata which informs decisions relating to the profile of an event.
- Next semester (Spring, 2017) I hope to collaborate with the 'i4inclusion' campaign to study the dynamics of hatespeech on social media.

For more information regarding cyberGIS analytics, see cybergis.illinois.edu
For more information regarding Hadoop, see hadoop.apache.org/docs/
USA Counties with FIPS and names.svg provided by commons.wikipedia.org
All logos are used under fair-use from their respective content holders

This project was made possible in part by my mentors: David Barbella, Xunfei Jiang, Charlie Peck, Dr. Shaowen Wang, Dr. Jiawei Han.