Title: Bridging the Gap: A Case Study of Integrating Social Media Big Data with Geography

Authors: Xiaokang Fu, Devika Kakkar, Jack Hayes

Keywords: Big data, High Performance Computing, Census data, Social media, Spatial Enrichment

Introduction

The fusion of social media data and geographical information presents a remarkable opportunity for real-time analysis of urban dynamics. Traditional Geographic Information Systems (GIS), face challenges in managing and processing such large-scale datasets. This paper aims to develop methods for the effective integration of big data with geography and demonstrate it with a case study. Our case study focuses on merging an extensive collection of 10 billion geotagged tweets with about 8.18 Million U.S. census blocks. Due to the large size of the dataset, our main challenge was to develop efficient methods to perform the spatial merge at this scale. By leveraging the capabilities of advanced geospatial data science, GPU-based databases, and High-Performance Compute (HPC) cluster we developed a novel approach to perform this merge in a time and cost-efficient manner. This innovative approach resulted in the creation of an extremely detailed social media dataset enriched with census information. The dataset offers unique geographic specificity, enabling us to generate insightful visualizations that depict social media trends and patterns across finely segmented geographic units ranging from country and states to counties. Political science researchers have used our resultant dataset to observe variations in real-time political expressions across granular geographic units for over a decade. Our dataset offers an unprecedented opportunity to observe public opinion with such temporal and geographic granularity. Our innovative approach makes a significant contribution to the field of political geography, demonstrating the potential of integrating social media data with geography for enhanced political analysis at a fine-grained geographic level.

Datasets

The project focuses on merging two valuable datasets of Goetweets and United States Census datasets. Both of these datasets are described in detail below:

The Geotweet Archive

The Harvard Center for Geographic Analysis (CGA) maintains a Geotweet Archive, a global record of geo-tagged tweets spanning time, geography, and language. The primary purpose of the Archive is to make a comprehensive collection of geo-located tweets available to the academic community. The Archive extends from 2010 to 2023. The number of tweets in the collection totals approximately 10 billion and is stored on Harvard University's High-Performance Computing (HPC) cluster. More information on the archive can be found here.

Administrative boundaries

Three administrative boundaries used in the project:

- State and County Boundaries: The US State and County Boundaries are from HeavyAI database's default table. The data source is from the U.S. Census Bureau. They contain ID, fips, county name, state name, and geometry fields.

- Census Blocks: We use the census blocks 2021 from the United States Census Bureau. These shapefiles contain multiple fields such as geometry, TRACTCE20 (Census Tract Code for 2020), BLOCKCE20 (Census Block Code for 2020.), GEOID20 (Geographic Identifier for 2020) and more.

Methodology

Our case study focuses on merging an extensive collection of 10 billion geotagged tweets with approximately 8.18 Million U.S. census blocks. Our major challenge was to develop methods to perform the spatial merge at this scale in a cost and time-efficient manner. By leveraging the capabilities of both GPU and CPU-based High-Performance Computing (HPC) clusters we developed a novel approach to perform this big data processing. Our workflow is shown in Figure 1 below:

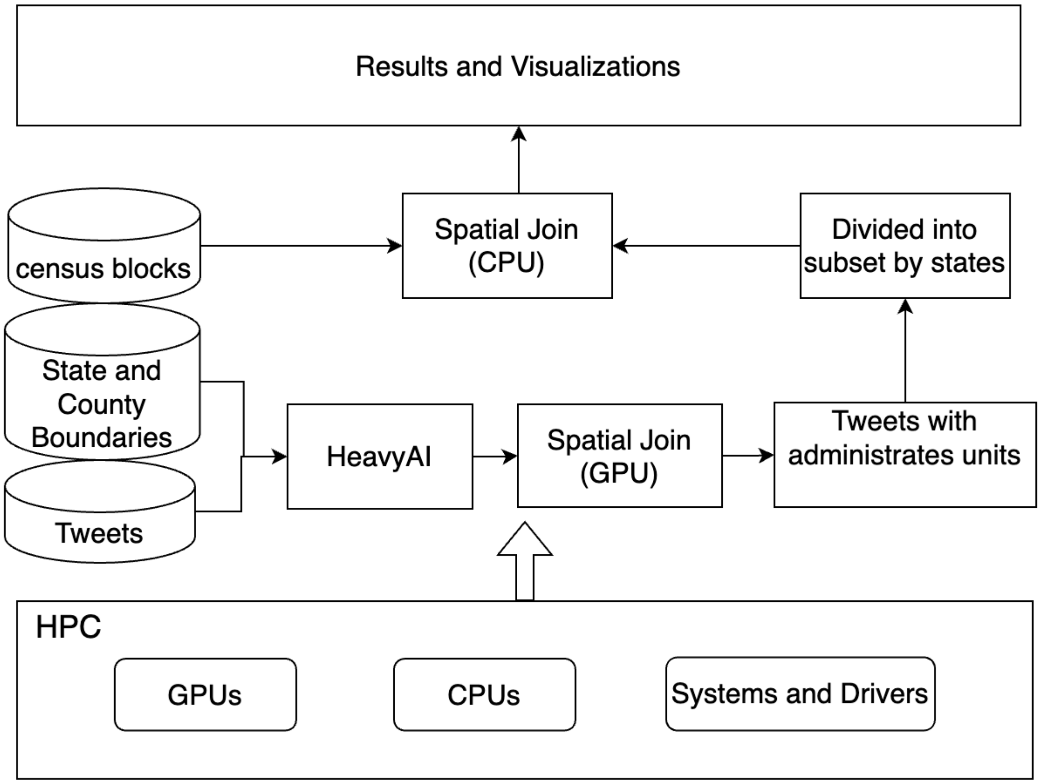
Results and Visualizations
census blocks
Spatial Join (CPU)
Divided into subset by states
State and County Boundaries
Tweets
HeavyAI
Spatial Join (GPU)
Tweets with administrates units
HPC
GPUs
CPUs
Systems and Drivers

Figure 1. Workflow to merge the geotagged tweets with administrative boundaries

It consists of a two-step solution as described below:

1. **GPU-based spatial enrichment:**

    a. We upload the daily geotagged tweet data to HeavyAI, a GPU-based database
    b. We use HeavyAI to do the spatial Join to join the tweets with state and county boundaries and store the result in the database
    c. We export the result and then delete the uploaded tweets and the results table to save the memory of the GPU
    d. We repeat steps a to c year by year and assign tasks to different nodes each year.

2. **CPU-based spatial enrichment:**

    a. We divide tweets into different subsets based on states and months of year. The state field is derived from the spatial enrichment done in step 1 and the month is derived from raw tweets. For each state, we do spatial join by performing parallel processing on the CPU to join the geo-tweets with census block information.
    b. We enrich every tweet with state and county boundaries and census block information, year by year, on different CPU nodes.

Results

Our work resulted in the creation of the most comprehensive archive of two billion tweets enriched with country, state, and census-level variables. This archive is available publicly for researchers on Harvard Dataverse as Harvard CGA Geotweet Census Archive. This archive is a subset of the Harvard CGA Geotweet Archive v2.0, enriched with nationwide census data. It contains the tweet and user identification records along with census variables for more than two billion geo-tagged tweets in the U.S. from January 2012 to July 2023. This dataset is available to the academic community at large, unlike the Harvard CGA Geotweet Archive v2.0 which is under Twitter's redistribution policy restriction for public sharing. To the best of our knowledge, it is the first dataset of geography-enriched tweets available at this scale and granularity.

The approach described here provides an efficient and affordable way to solve geospatial big data problems particularly those involving spatial merge and enrichment. It is cost and time-effective with extremely fast Input/Output speed. For example, using our two-step approach, we can enrich 10 Billion tweets with 8.18 Million census blocks in just 1.5 days. The GPU-based enrichment of country and state-level variables took 8 hours and the census block enrichment took about 1 day. For the GPU-based enrichment on Heavy.ai, we used 13 A100 GPUs. For the CPU-based enrichment, we used 13 CPU nodes with 60 cores, and 480 GB RAM for each node. Our source code is open-source and available publicly on GitHub.

Our resultant dataset is currently being used by political scientists to examine the impact of geography on political opinions expressed on social media across increasingly small geographies such as census blocks. Further, we utilize this data to conduct a comprehensive sentiment analysis, investigating how emotional tones and public opinions vary across space and time. This involved examining the nuances in positive, negative, and neutral sentiments, enabling a deeper understanding of regional and temporal differences in political attitudes and communication styles.

Conclusion and Future Work

The integration of big data with geographic analysis, as demonstrated in our study, has far-reaching implications for political science. It opens new avenues for researchers to observe real-time political expressions across granular geographic units over time. Compared to traditional GIS methods, our approach offers significantly enhanced speed, accuracy, and granularity. While traditional methods struggle with the scale of social media big data, our methodology demonstrates a feasible solution to this challenge. This research represents a significant leap in the field of geospatial big data analytics. By successfully integrating a massive dataset of geotagged tweets with U.S. census blocks, we have demonstrated the potential of combining such big data with geographic analysis and highlighted its impact with a case study.

Future research could explore the application of our methodology to other types of datasets such as WebAI data, and climate change data. Additionally, further refinement of data processing techniques could enhance the efficiency and applicability of our approach. In addition, we also plan to explore ESRI geo-analytical servers for performing this task and compare it with the method used in this paper.

Acknowledgment

We extend our deepest gratitude to the National Science Foundation (NSF) for their generous support of our research. This work was made possible through NSF award #1841403, which provided critical funding and resources essential to the development and success of our project.