Twitter Map Reduce Methods and Results

[Extending the cyberGIS toolkit]

Ben Liebersohn
Earlham College
801 National Road West
Richmond Indiana
btliebe12@earlham.edu

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 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 punlished annually, sifting through such an enormous database has proved nontrivial.

Though database size reduction, this Hadoop based 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 studying 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.

Categories and Subject Descriptors

Datamining [**Information Systems Applications**]: Twitter; High Performance Computing [**Hadoop**]

General Terms

Big Data

Keywords

cyberGIS, Twitter

1. INTRODUCTION

One of the challenges facing this project has been a hurdle faced by many computational and theoretical Geographers before me: topography. Spaces exist in many ways topographically, and these topographic spaces change much faster than geographic boundaries.

Topographic data extracted from the Twitter streaming API, such as locations, and bounding coordinates, can tell users about the behaviour of users who submit content on social media platforms. The data collected using the Twitter streaming API on the ROGER supercomputer is collected with geographic intensive computation and analysis in mind.

As part of the XSEDE network, ROGER is the cyberGIS supercomputer, as it provides a heterogeneous computing model for geographic analysis. While this limits some of the potential computability of certain kinds of machine learning algorithms which may run better on a homogeneous computer, mapping vectors and coordinates is made simple on ROGER with its inclusion of GPGPU components.

The CyberGIS Center at the National Center for Supercomputing Applications has been collecting data from the Twitter Streaming API since January 2013, a catalogue of data which spans 3 years. This allows me to explore current phenomena with a small degree of historical context. This dataset will be looking at users in the United States, from which the Streaming API provides 1 percent of the total throughput of traffic on Twitter.

Two examples I will discuss as potential use cases will show how a number of topics ranging from sports to politics gives an insight into how physical spaces related to their cyber counterparts.

First, we will be looking at American Football, and understanding how trends of communication regarding football may shift depending on who is winning, and where that team is from. Secondly, we will look at the 2016 election, and try to understand what the relative popularity of candidates was at different times in cyberspace, trends which we can compare to polling at different times.

The Twitter data collected on the ROGER Supercomputer is directly files into an HDFS database for examination using Hadoop. The Hadoop framework is great for mapreduce style operations. This will work for some important features, but not all of them. One thing Hadoop will primarily do for me is 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. In my conclusion, we will talk about getting the data ready for PostgresSQL, a task which will hopefully take place next semester. This will be computed on ROGER, much like the Hadoop experiments described in this paper, however instead of a dedicated node I will use

I am able to leverage our local PostgreSQL users whose technical insights will be valuable in future exploration of our data. From a relational database, I am more able to query for the spatial and temporal themes. This is especially true for PostgreSQL database, which includes PostGIS and allows me to parse geospatial information quickly. My schema relied mostly on the spatiotemporal analysis, in which Combined with python data analytics tools, my project seeks to classify spatiotemporal events, with the goal of ultimately generating metadata in reference to the context in which events take place.

2. RELATED WORK

My project is guided towards the future potential uses. Specifically, by finding ways in which we leverage the classes of events highlighted on Twitter, it will be possible to respond to events in a distributed information hub. First responders may be able to respond to more events 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 platform for detecting disaster.

Timescales of events not only unifies 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. Existing methods such as the scalable spatiotemporal analysis method GeoBurst help identify events spatiotemporally. Combined with additional contextual decision making, an event classifier can become a dynamic tool for understanding communication patterns.

"CatchTartan: Representing and Summarizing Dynamic Multicontextual Behaviors" [2]

By using a multicontextual, dynamic organization of keywords, the data input can be organized in different ways, including temporal clustering. This can be applied to behavioural data, which may be inconsistent enough to warrant a multidimensional approach to learning and processing. This is helpful for understanding Twitter data, but the use cases for such a technology are very broad. It might be worth investigating other uses for this data than the ones presented in the article.

–

"GeoBurst: Real-time Local Event Detection in Geo-tagged Tweet Stream" [3]

This paper describes a method named âĂŸGeoBurstâĂŹ which has the ability to detect correlated geo-tagged tweets. As patterns emerge, events are proposed, the idea being that this can find trends which are spatially connected. This article proposes a method which is faster and more accurate than previous methods. This can provide a platform for live analysis, but also fast and accurate weak scaling. The purpose of seeking a weak scalability approach would be to contrast a live stream of data against a historic growing dataset.

–

"Scalable Topical Phrase Mining from Text Corpora" [4]

By using a different method for finding keywords and associated constructions which a body of text associates. This paper widened my perspectives on how multicontextual behaviours are not only dynamic and complicated, but also very tracable and interoperable. This method is supposedly scalable and faster, though it is unclear how well it can network phrases in multiple dimensions. Instead, what I took away from it was that instead of looking at only one kernel phrase, it weighs the connections as polynomial representations of phrases. This could be applied to Tweets to broaden our search vocabulary in our massive Twitter history.

–

"Mining Multi-Aspect Reflection of News Events in Twitter: Discovery, Linking and Presentation" [5]

Similarly using a dataless classification scheme (which would be useful for my live data), this paper promotes âĂIJa unified framework to mine multi-aspect reflections of news events in Twitter.âĂİ This paper contains specific queried insights into live data Twitter API streams. This method is effective at selecting meaningful topics from the âĂŸnoiseâĂŹ of Twitter.

–

"GIN: A Clustering Model for Capturing Dual Heterogeneity in Networked Data" [6]

After reading this, I saw that this article seeks to encapsulate as much information as possible in its representations of the models of information pathways in networks. Rather than using traditional node-link cluster representation, the methods used allow for efficient and scalable EM classification. The purpose of this for my project would be as a starting place for encapsulating the live and historic data sets.

3. DESIGN

Spaces exist in many ways geographically. Topographic data extracted from my datasets, such as locations and bounding coordinates with various granularity can tell us about the behaviour of the users who submit the content. Sports fans are often determined geographically via proximity, sometimes transcending features like State borders and instead giving cultural regions unique boundaries. This is an important consideration with regard to how my sports analysis is conducted. Political boundaries are often understood through polling in the electoral college. Making thematic maps may give us new insights into how the users of social media reflect the choices of the electorate in that area. For these reasons

it is unclear how representative Twitter is of the general electorate.

Because this project could be applied to a wide range of topics, I chose two topic queries which are massive phenomena in social media, while connecting geospatial regions in time. Past experimentation with geotagged tweets relates largely to sentiment analysis. As this field of AI continues to develop, Twitter will certainly continue to be one focus of this research. Sentiment during many temporal events may vary, and not only can this tell us about contemporary events, but it will also serve to form a historiographic context for future analysis of events.

Currently, much of the interest in Twitter data mining is to advertise products and find optimal audiences for capitalist venture. While this is not the primary goal of my project, however it is noteworthy as it provides a downstream utility for most of my work. The effectiveness of this is likely there, however it is unknown how representative these datasets are for consumers. In hindsight, trends may be apparent, however the competitive consumer may not consider the Twitter space as representative of their views.

My first topic, sports, will rely mostly on the conversations which discuss the individual teams, players, and games. By looking for sports related messages, one can sample the time before during and after the events. Because of their predictable frequency and broad appeal, a clear delineation of sports fan-bases emerges, as social boundaries related to geographic proximity tie locations to favouring events which feature nearby teams. Continuing on the work of GeoBurst, I believe this program may better inform cultural boundaries which relate to proximity to unique urban zones. [3]

Studying the relationship between politics and social media is one goal of this project. Social media campaigns ranging from Bernie SanderâĂŹs grassroots organizing, to likely outside interference from Russia and itâĂŹs social media presence are all important to study. This is not only in light of known cyber-activist political campaigns but also in the effort to uncover new methods which will continue to be utilized as long as they are viable. The Hadoop code which runs on ROGER is designed to select data from an HDFS file system based on the time events, spatial events and more. By filtering my data set on these parameters, it becomes manageable in a PostgresSQL database.

The filtering of this data takes place within the map reduce stage of my project. Taking in data from the Twitter Streaming API, I reformat it first to the HDFS partition. HDFS (Hadoop Distributed File System) is designed in large part for scalability and load balancing on clusters. This process begins with breaking up groups of tweets based on chronological day date system, which is saved into seperate files.

Once copied from local HDFS to each node on the cluster, Tweets are individually mapped to the boundaries or spaces I want to look at. From there, we begin the reduce stage. This begins by taking out Tweets which do not fit our timescale. From there, removal begins using clipping. If each of the GPS boundaries fits into the user specified domain, it is finally copied back to local to a new file.

This design allows for more flexibility in how map reduce is implemented because it uses Hadoop for the map reduce stage. Once logged into a Hadoop node on ROGER, I instruct Hadoop on how to run the job. This is usually done either as a interactive job or run in a bash script. This allows for more debugging and testing than just scripts, as we can see the procedural results.

The user first supplies the local memory space we are using, as well as the upper bound on the number of reducers (Hadoop has the final say on this) as well as a time range to look at. From there we continue by giving coordinates which bound the spaces we are looking at. The data is copied into a new location (so as to preserve the original in a generic filesystem for the cluster to break up.

Then we instruct Hadoop to use the mapreduce code I wrote in Java. This code first maps the points in each message from the correct time range. From there, the reducers perform a clipping operation. Messages which go unclipped are copied to a new generic filesystem.

From there, the resulting output is recursively copied into an output file which contains the relevant messages.

As much as sports teams may represent a city, so too does political discussion. Social media has been influential in politics, though that relationship has yet to be firmly connected to geographic traits. In this unprecedented election cycle, one valuable topic is the Donald Trump social media phenomenon, and a more in-depth review of the role of hate. Numerous forms of cyber-hate include trolling, harassment, and crimes made possible through social media. With a pipeline such as CatchTartan [2] and text mining [4] it is possible to trace these issues which exist in both real life and digital spaces in dense urban zones. These dynamics also play out in many settings through centralized influences, and finding the qualities which influence different groups may play a key role in undermining hatespeech and instead promote inclusive behaviours.

4. RESULTS

Twitter data comes in many formats, in part because of the changing trends among users. Users are not to blame for these changes in trends however, as many features have changes, and the interface for communication on social media (topics discussed, forms of media presented, etc) have changed greatly from 2013 until now. This has impact in other areas as well, with the focus of many geospatial data being a form of 'check in' from local businesses or other locales. This category, while designed to promote spending and in effect become a form of advertising, is much harder to parse spatially, but might be an interesting topic to look into.

The ability to select spatial zones which are of potential interest limits the geospatial contextual scope of a dataset. This is important for event detection as it leads to fewer false-positives and unrelated topics. This feature was more difficult to impliment than others because it relied on an interdimensional approach wherin I merged many different

patterns and behaviours in a concise manner. While this is critical, another feature which is of greater significance is the ability to chose a timescale. Most of the regions we are looking at are already dense in terms of activity, so more importantly we must decide what timescale we are interested in investigating. The spatial connectivity of social media is very high, allowing far away locales to interact. For this reason, looking at a specific scope of time will later provide the temporal burstyness needed to find using data mining methods such as GeoBurst.

Much of the best results come from filesize reduction and the relative limits on such abilities. 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. Continuing from there, if one were to reduce the data to Tweets mapped in Indiana and reduced to 12/11/2016, I found that they amounted to only 28MB, which is very impressive. This gives us metrics for computable of Twitter messages which may interest social scientists looking for individual events.

Features such as geo-tagging, a particular concern of ours have also changed. Not only does Twitter now provide locale names, it may instead print GPS coordinates which bound the locale, rather than using a singular coordinate. This has increased the body of geospatial data, however it has decreased the value of each individual message, as many include boundaries which are too broad to include in an output.

An example of a Tweet which gets thrown out is: [message and username removed] 288de3df481163e8‘ Alabama‘Alabama, USA‘ United States‘ Admin‘ Null‘

(30.144425,-88.473228),(35.008029,-88.473228),

(35.008029,-84.888247),(30.144425,-84.888247),‘ Polygon‘

https://api.twitter.com/1.1/geo/id/288de3df481163e8.json‘

Alabama, USA‘ 337‘ 36‘ 72‘ Sun Sep 30 21:05:13 CDT 2012

Rather than keep such a message, we may instead note it is too large of a bounding box to provide meaningful details. For much smaller bounding boxes in dense urban zones, it may be more advantageous in the future to find a centroid which defines the spatial approximation of the wherabouts of a user.

5. CONCLUSION AND FUTURE WORK

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. This pays a distinct role in decision making for many groups. Among those groups is performers in search of audiences, products looking for stores, logistical operators trying to localize popular trends early, and many others. This clear link to marketing makes social media topography an important topic for many laypersons whose interest in data is largely in the Twitterverse

Filesize is approximately days * density / average USA tweet density * 1.5GB assuming 200,000,000 messages per day. This means that on a low traffic year (as it is rare for there to be under 330,000,000 messages per day) upwards of a half billion Tweets, or almost 100TB of raw text is being produced. Combined with a larger database, the Hadoop module will be able to carry out the strong scaling required for massive calculations on more cores in more advanced environments. Many of the mapping procedures will scale well for mapping on GPGPU architecture. This can also possibly translate into a CUDA reworking of the map stage, with reduce being handled in the same pipeline.

Topographic data extracted from my datasets, can tell us about the behavior of the users who submit Tweets in a specified spatiotemporal zone. While some percentage is thrown out, the highest engagement is found among well defined spatial zones. This is possibly due to a link between the locale and the liklihood that a person is more active and surrounded by active users. In other words, a New Yorker may not be concerned about the privacy of their locale due to the density of human activity to begin with. Knowing one is in Harlem doesn't give them away in a way which might make rural users uncomfortable. Another possible connection is the age dynamics relating to urban Twitter usage and sharing of locales in more messages.

This method of Tweet extraction is a valuable part of cyberGIS data analytics toolkit. While small, it provides a pipeline for users and collaborators with the ROGER supercomputer to begin working with more social media data than they would otherwise be able to run on local machines. One goal of the CyberGIS community is to reach out to other disciplines. As an interdisciplinary field whose data richness is used by many academics, it is in part the job of the existing community to make their toolkit available to others who may need them. Part of this means also talking to social scientists, academics whose interests in computational work may be limited. Their role as analysts is important, and it is crucial that we do not interpret beyond our domain. *Sutor, ne ultra crepidam.*

Next semester (Spring, 2017) I hope to collaborate with the 'i4inclusion' campaign to study the dynamics of hatespeech on social media. From conversations with members of this group, and its convener Welling Hall, there is an interest in the idea of studying hate on the social media platforms used by not only activists, but also political organizations. Both hate and inclusion are interests in the i4inclusion project, which looks for discriminatory tactics and beliefs perpetuated by the media. By looking at trends in text keywords, user responses, links, hashtags and perhaps images, bursts of activity which are linked to the

By filtering my dataset on spatiotemporal parameters, s by consolidating reports with Twitter. In addition, bottlenecks on dense urban communication networks is eased in timesof 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.

6. BIBLIOGRAPHY

–[1]

Liebersohn, Benjamin "Annotated Bibliography", Earlham College Senior Capstone, Oct. 7, 2015.

–[2]

Meng Jiang, Christos Faloutsos, Jiawei Han, "CatchTartan: Representing and Summarizing Dynamic Multicontextual Behaviors", in Proc. of 2016 ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (KDD'16), San Francisco, CA, Aug. 2016

–[3]

Chao Zhang, Guangyu Zhou, Quan Yuan, Honglei Zhuang, Yu Zheng, Lance Kaplan, Shaowen Wang, Jiawei Han, "GeoBurst: Real-time Local Event Detection in Geo-tagged Tweet Stream", in Proc. of 2016 ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'16), Pisa, Italy, July 2016

–[4]

Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, and Jiawei Han, "Scalable Topical Phrase Mining from Text Corpora", PVLDB 8(3): 305 - 316, 2015. Also, in Proc. 2015 Int. Conf. on Very Large Data Bases (VLDB'15), Kohala Coast, Hawaii, Sept. 2015.

–[5]

Jingjing Wang, Wenzhu Tong, Hongkun Yu, Min Li, Xiuli Ma, Haoyan Cai, Tim Hanratty, and Jiawei Han, "Mining Multi-Aspect Reflection of News Events in Twitter: Discovery, Linking and Presentation", in Proc. of 2015 IEEE Int. Conf. on Data Mining (ICDM'15), Atlantic City, NJ, Nov. 2015

–[6]

Jialu Liu, Chi Wang, Jing Gao, Quanquan Gu, Charu Aggarwal, Lance Kaplan, and Jiawei Han, " GIN: A Clustering Model for Capturing Dual Heterogeneity in Networked Data", in Proc. of 2015 SIAM Int. Conf. on Data Mining (SDM'15), Vancouver, Canada, Apr. 2015 (selected as one of the best papers in the conference and invited to journal Statistical Analysis and Data Mining (SADM) special issue "Best of SDM 2015")