Demonstrating PlanetSense: Gathering Geo-Spatial Intelligence from Crowd-sourced and Social-media Data

Gautam S. Thakur, Kevin Sparks, Roger Li, Robert N. Stewart, Marie L. Urban

The Geographic Information Science and Technology (GIST) Group

Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37831 USA

{thakurg, sparksa, lirr, stewartn, urbanml}@ornl.gov

ABSTRACT

Crowd-sourced and volunteered information, social media, and participatory sensors are capable of providing real-time activity data. Monitoring these sources in time of relevance and then using them to gather operational knowledge is important during crisis management. Beyond that, it's important to curate this information for geo-spatial research purposes, including land use classification and population occupancy analysis. In this demonstration, we will showcase PlanetSense - a geo-spatial research platform built to harness the existing power of archived data and add to that, the dynamics of heterogeneous real-time streaming data from social media and volunteered sources, seamlessly integrated with sophisticated machine learning algorithms and visualization tools. A demonstration will focus on – 1) Recent initiative emphasizing the need to harness crowd-sources, volunteered, and social media data at scale; 2) Anatomy and insight into data collection workflow. We will show the ability to harvest and process several terabytes of raw data in real-time; 3) A detailed discussion with insight into more than 20 sources of data will be given. These sources include text, sensors, as well as imagery data; 4) PlanetSense's end to end distributed architecture will be discussed with focus on collecting and processing high-volumes of streaming data in a Geo-Data Cloud. Data fusion methods and algorithms for integrating disparate data sources with existing legacy products. Data analytics and machine learning methods for generating operational intelligence on the fly; 5) In addition, PlanetSense "App" platform will be shown with hands-on application enabling interested audience to quickly develop and deploy solutions. 6) Several case studies will be discussed relevant to, land use classification, monitoring transient population, high-resolution occupancy analysis, mapping special events population, ability to uncover global breaking events and reactions in near-real time, ability to track protest, unrest, and monitor other societal turbulences as they happen, and real-time monitoring of infrastructure outages.

CCS Concepts

• Information systems ~ Geographic information systems • Software and its engineering ~ Data flow architectures

Keywords

Social media, volunteered geographic information, big data

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.

Copyright is held by the owner/author(s).

SIGSPATIAL'16, October 31 - November 03, 2016, Burlingame, CA, USA
ACM 978-1-4503-4589-7/16/10.

<http://dx.doi.org/10.1145/2996913.2996975>

architectures, real-time analytics.

1. INTRODUCTION

Recently, social networking services can provide enormous amounts of information on people's diurnal activities, personal preferences, relationships, personal/political/professional views, and inter-personal behavior. This information is accumulated in form of rich texts, images, and videos across space and time over a long period. Key insights discovered after combining these sources is vital for the community and to first responders during emergency and natural disasters. In particular, annotated information and text data from a region helps to capture any activity and to uniquely fingerprint each location within. It might also be used developing new models for land use classification and occupancy analysis[XXX]. However, this type of data is difficult to harvest at large-scale and even more difficult is to extract meaningful insights from them[XX]. Our previous experience has shown that traditional computing architectures relying on relational data-stores are incapable of harvesting multivariate and multidimensional data (both text and imagery); while conventional machine learning approaches are inadequate in processing with high accuracy. In this work, we address all these issues, first we propose PlanetSense – An end-to-end scalable platform that is purposefully designed to harvest, analyze, and visualize massive amounts of crowd, social, and open source imagery and other textual data; simply, quickly, reliably, for gathering "real-time" geo-spatial intelligence. Second, we demonstrate this unique capability by applying methods to identify and classify information from several billion records, gathered from crowd-source and social media outlets. Finally, to intuitively interact and understand, a dynamic visualization tool will be shown as an integrated PlanetSense module.

2. ARCHITECTURE

In this section, we briefly describe the big data computing architecture supporting our need to perform data collection and analysis. Figure 1 shows PlanetSense architecture and its constituent components. A scalable architecture is developed and deployed on leadership class computing infrastructure[1]. This end to end system provides a set of coherent services through its Geo Data Cloud, Metis ML Server, and Elastic Compute cluster. It is designed to harvest, analyze, and visualize massive amounts of crowd, social, and open source data simply, quickly, reliably, for geo-spatial research purposes. In this work, we will showcase how this platform allowed us to perform spatial analysis for land use classification into facility types and capturing transient population. In-depth information on this platform is available at [2]. Next, we briefly explain the architecture-

Figure 1: PlanetSense architecture

2.1 Geo Data Cloud

The Geo Data Cloud is a purpose-driven data collection cluster, harvesting real-time information from major social media, news, and events channels. A compendium of more than 20 sources with thousands of variables spread across images, sensor, and text data.

2.2 Elastic Compute and Real-time Streaming

This forms a central part of the platform allowing users to compute the result of their query in real-time by adaptively tapping into its given pool of resources. The purpose of real-time streaming layer is to compensate for GeoData Cloud's access latency and lagged processing time. It has a small data footprint and works by first caching the streaming data entering from outside sources and then processing minimally, generating real-time information.

2.3 Metis ML Server

A machine learning server cluster for prediction analysis and contains a scalable data processing and analytics layer. The algorithms consume the processed information (archived data such as census, LandScan) and streaming layer (e.g. Twitter Streaming) into the merged layer. The merged layer also stores trained machine learning models such as topic models and SVM classifiers that extract scenarios based on data collected through searches. The visualization library complements analytics and serves as a high-performance 2D/3D platform.

2.4 Presentation Layer

The presentation layer provides multiple interfaces to interact and gather geospatial data on the fly. There are currently three different interfaces provided based on the kind of operation required for the information generation.

2.4.1 Export Dataset

Users wanting to store and use the data can download the processed information containing the relevant research.

Figure 2: PlanetSense core capabilities

2.4.2 Web Interface

The web interface is designed to provide inference based learning extracted from the analytics layer. Also, the results of the algorithms are modeled and plotted through the interface layer.

2.4.3 Query Interface

This is an interactive interface that provides ad hoc execution of user queries. This interface provides succinct information in the form of output to end users and third party products.

3. PLANETSENSE CORE CAPABILITIES

PlanetSense is built to serve multiple research interest from exploring relevant information to generate operational intelligence. In Figure 2, we have shown platform's core-capabilities.

3.1 ActivInt

The platform can be used for pursuing activity based intelligence using social media data. Activity Based Intelligence (ABI) is an algorithmic driven methodology to process and extract insight about certain activities from the data. The idea behind this approach is focus on patterns and their propagation.

3.2 Explorer

This module provides insights into exploratory nature of finding things from multitude of social media contents. Some of specialties that are provided includes social statistics, keywords searching across multiple data sources, and provide insight into certain areas of interests.

Figure 3: Museum occupancy

3.3 Diutius

This module enables user to perform spatio-temporal analysis as well longitudinal analysis on data patterns and the curated content in PlanetSense platform. One of the studies that we pursued was to examine facility occupancy such as museums, sport venues, etc. Figure 3 shows the distribution of museum goers based on social media data analysis. This involves analyzing museums goers' check-in patterns for period of over 90 days.

3.4 GeoQuest

Geo-location services are becoming prevalent for several reasons. Some of them includes providing friend recommendation, advertising, and a host of services for tourist and visitors. This module enables exploring multiple social media data sources extracted from given geo point or bounding box. The aggregated information provides insight into multifaceted insight through text, sensor data, and images.

3.5 NeXus

NeXus provides linked-data analytics to discover links between multiple sources. Some of the examples includes, links between energy-water, carbon emission – pollution. This module is built on customized graph libraries that provide insight into fast searching based on annotated links (Edges of a graph).

3.6 TipNCue

Tipping and Cueing enables noticing something interesting is going on and directing further insight to explore at that event. This module works in a feedback loop fashion and therefore store queries across multiple sessions and requests to create better information about an entity.

4. TRANSIENT POPULATION DYNAMICS

For past several decades, research efforts in population modelling has proven its efficacy in understanding the basic information about residential and commercial areas, as well as for the purposes of planning, development and improvement of the community as an eco-system. More or less, such efforts assume static nature of population distribution, in turn limited by the current ability to capture the dynamics of population change at a finer resolution of space and time. Fast forward today, more and more people are becoming mobile, traveling across borders impacting the nuts and bolts of our urban fabric. Unfortunately, our current efforts are being surpassed by the need to capture such transient population. It is becoming imperative to identify and

Figure 4: Transient population dynamics

define them, as well as measure their dynamics and interconnectedness. Using the platform, we intend to research urban population mobility patterns, gauge their transient nature, and extend our knowledge of their visited locations. We plan to achieve this by demonstrating novel method and using VGI data that models and characterizes transient population dynamics. Earlier, we developed a transient population model that is illustrated in [3]. We used this model and data from several social media sources to estimate transient population and discover locations (facilities) in Australia. We compare our aggregate results with the census of Australia in Figure 4. We have circled out regions containing transient population, that is missing on the census map. This approach can be very helpful in uncovering regions which are otherwise not represented using traditional techniques (e.g. census and surveys).

5. LAND USE CLASSIFICATION AND OCCUPANCY ESTIMATION

The use of open source and crowd contributed data has tremendous ability to provide insight into occupancy analysis for small area estimates such as building and airports[4]. A longitudinal study can provide insight into spatio-temporal occupancy of these areas that are important for settlement analysis. However, access to such a fine-resolution is data is very difficult and collecting such data has scalability challenges. In this work, we explore harvesting social media data from Facebook and Twitter to augment our analysis of occupancy at scale. We have used social media check-ins and likes to discover occupancy and facilities categories. We have monitored the changes and collected changes in check-ins and likes for some time. Then we have deployed Linear SVM models to classify more than 700 Million tweets into facility types such as restaurants, airports, and sports avenues[5], [6]. The result of our analysis from Australia and United Kingdom data are shown in Figure 5 and Figure 6 respectively. The analysis indicates that in United Kingdom,

Figure 5: Australia facility counts

Figure 6: United Kingdom facility counts

‘Shopping and Retail’ as well ‘Professional Service’ facilities exceed any other facility types. They combined for nearly 20% of all businesses found. Also, ‘Spas Beauty Personal Care’ found to be nearly similar to these facility types. Sports and Recreation is a distant fourth among other facility types. Australian facilities are quite restricted (less than 70), however the overall distribution of facilities is more even than other countries. Similar to United Kingdom, ‘Shopping and Retail’, ‘Restaurants and Cafe’, ‘Professional Services’ have a larger presence in Australia. However, ‘Sports Venue’ seems to be more prominent than other facilities and provides a unique signature for this country.

6. CONCLUSION

In this demonstration, we will showcase our big data capability to analyze several billions of social media and VGI data points to generate geo-spatial insights. We will provide insight into the design and development of PlanetSense infrastructure and its ability to ingest multi-terabyte data through elastic infrastructure. Using case-studies, we will show the efficacy of using social media data in generating transient population signatures, land use classification, and occupancy analysis. In the future, we would like to open this platform to the public, particularly for students and academics to aid in their research.

7. ACKNOWLEDGMENTS

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05- 00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up,

irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

8. REFERENCES

- [1] O. R. N. Laboratory, “The Oak Ridge Leadership Computing Facility (OLCF),” 2015.
- [2] G. S. Thakur, B. L. Bhaduri, J. O. Piburn, K. M. Sims, R. N. Stewart, and M. L. Urban, “PlanetSense: A Real-time Streaming and Spatio-temporal Analytics Platform for Gathering Geo-spatial Intelligence from Open Source Data,” Jul. 2015.
- [3] U. M. Thakur Gautam, Sparks Kevin, Stewart Robert, “Curating Transient Population in Urban Dynamics System,” in *Arxiv*, 2016, vol. 1, no. 1, p. 6.
- [4] R. Stewart, J. Piburn, M. Urban, and A. Morton, “Can social media play a role in developing building occupancy curves for small area estimation?,” *geocomputation.org*.
- [5] R. Stewart, M. Urban, S. Duchscherer, J. Kaufman, A. Morton, G. Thakur, J. Piburn, and J. Moehl, “A Bayesian machine learning model for estimating building occupancy from open source data,” *Nat. Hazards*, vol. 81, no. 3, pp. 1929–1956, Jan. 2016.
- [6] I. H. Witten, E. Frank, and M. A. Hall, “Data Mining: Practical Machine Learning Tools and Techniques,” Jan. 2011.