SAND2019-1805C

Turning Big Data into Actionable Intelligence

Forest Danford, fdanfor@sandia.gov, (505)844-1186
Laura Patrizi, lapatri@sandia.gov, (505) 284- 2427
Sandia National Laboratories, Albuquerque, NM 87185-0810

ABSTRACT

At Sandia National Laboratories (Sandia), we are interested in quickly turning near-real-time streaming data from a variety of sources into actionable insights that enable decision makers, from political leaders to field commanders, to take appropriate actions when faced with security threats. The SNL “Turning Big Data into Actionable Intelligence” project is focused on understanding, integrating, and evaluating technologies that make it possible to quickly discover actionable intelligence from geospatially-based, large volume, and disparate data sources.

BACKGROUND

We live in a world experiencing a monumental increase in the amount of available digital information. “90% of the data in the world today has been created in the last two years alone, at 2.5 quintillion bytes a day” – IBM Marketing Cloud (December 2016). Human synthesis of data at scale is impossible. Big data tools, including machine learning, are useful for highlighting significant information for human decision makers, especially in time critical situations. Sandia National Laboratories (SNL) is interested in quickly turning near-real-time streaming data from a variety of sources into actionable insights that enable decision makers, from political leaders to field commanders, to take appropriate actions when faced with security threats.

PROJECT DESCRIPTIONS

The SNL “Turning Big Data into Actionable Intelligence” project is focused on understanding, integrating, and evaluating technologies that make it possible to quickly discover actionable intelligence from geospatially-based, large volume, and disparate data sources. With our partners at the University of Illinois and the Applied Research Institute (ARI) of Chicago, we have developed an architecture of automated data processing pipelines on an open source software stack. This architecture allows mapping of multiple heterogeneous data source (structure and unstructured) into a common data schema. This design provides a framework for machine learning algorithms to learn and analyze data (e.g. near real-time analytics) from heterogenous data sources and turn them into actionable information for decision makers. Our architecture is scalable to support increases of data sources

Architecture tools include the Hortonworks Data Platform, Zookeeper and Ambari for process coordinating, Kafka and Storm for data streaming, Solr for data search, TensorFlow for machine learning. Lucidworks Solr Banana is used for data visualization. Primary components of a data processing pipeline are: a data getter for acquiring the data, a normalizer enabling meaningful comparisons, a processor performing data type specific analysis (e.g. automated object recognition), and a publisher making acquired data and analysis results available to further processing and human analysists. A merged neural network is combined with these pipelines to discover insightful information.

To vet our architecture concept, we selected an exemplar problem with readily available geospatially-based data: near-real-time traffic estimation. Due to our partnership with ARI, we focused on Chicago traffic data. Our goal is to utilize a variety of data, both images and text, from multiple data sources to provide a Chicago

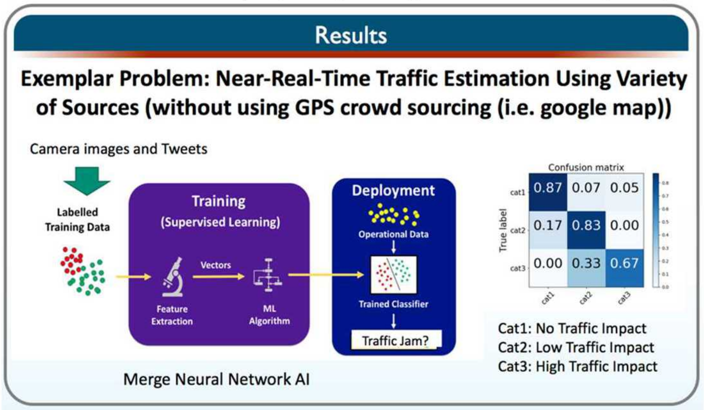
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy’s National Nuclear Security Administration under contract DE-NA0003525.

traveler with route planning guidance. We developed and deployed data processing pipelines on SNL dedicated data analytics and machine learning hardware clusters. Data being processed include: Twitter traffic related tweets; City of Chicago traffic data; MapQuest incident reports; TravelMidwest web camera images, vehicle detection system readings, and dynamic message signs; Digital Globe satellite imagery; and Dash Camera Video. We process around 300K image and text records per day.
As an example of how a mixture of these data types could provide insight, consider a scenario where the following records, coincident in both time and location, are found:

- Tweet: “omg traffic is bad” (geo-tagged), or “avoid Bishop freeway” (georeferenced)
- MapQuest Traffic incident report: “Delays increasing, and delays of six minutes on Bishop Ford Freeway, Average speed 15 mph”
- Web camera image: containing many automatically identified cars

From these records, a Chicago traveler may conclude that there is a traffic jam on Bishop Ford Freeway. This is obviously an ideal scenario. It is not, however, an unlikely one if bounded by a reasonable time interval (e.g. 15 minutes), and geospatial region (e.g. 1 mile).

Our data processing architecture allows for these types of data records to be analyzed by our merged neural network to produce indications of traffic levels (no, low, or high traffic impact) at queryable locations.



Results

Exemplar Problem: Near-Real-Time Traffic Estimation Using Variety of Sources (without using GPS crowd sourcing (i.e. google map))

	cat1	cat2	cat3
cat1	0.87	0.07	0.05
cat2	0.17	0.83	0.00
cat3	0.00	0.33	0.67

Cat1: No Traffic Impact
Cat2: Low Traffic Impact
Cat3: High Traffic Impact

Some of the significant technical challenges encountered during the development of this project:

- mapping all data types to a simple enough schema that enables timely machine data synthesis for creation of a human-interpretable, coherent decision
- acquisition of enough information to train the merged neural network
- generic code base creation to enable quick development of new data pipelines
- long-term maintenance of data pipelines (intermittent services, computers restarts, etc.)

Future work will focus on the challenge of providing meaningful uncertainty quantification of insights.

BIOGRAPHIES

Laura Patrizi has an M.S. in Computer Science from the University of New Mexico, and a B.S. in Computer Science and Mathematics from Colorado State University. She has over 13 years of experience with satellite ground station software systems. For the past 3 years she has been working as a software engineer at Sandia National Laboratories on big data processing projects.

Forest Danford has an M.S. in Computer Science, a B.S. in Computer Science, and a B.S. in Biosystems engineering from the University of Arizona. He has been a software engineer for the last 5 years at Sandia National Laboratories. He enjoys big data problems, data visualization, and developing analytics to serve the national interest.