Geospatial Big Data for Finding Useful Insights from Machine Data

Pouria Amirian*, Francois Van Loggerenberg, Trudie Lang and Margaret Varga

The Global Health Network, The University of Oxford

November 6, 2014

Summary

The focus of this paper is on finding useful insights from data generated by Point-Of-Care (POC) Diagnostics devices based on spatio-temporal data analytics. At first glance data generated by the molecular POC diagnostic devices seems not very relevant to geospatial Big Data. However by including location of the devices (which are usually locations of the healthcare settings e.g. hospitals or labs) in the analysis and by inclusion of other geographic layers, whole new set of useful questions can be asked and therefore many useful insights can be found.

KEYWORDS: Geospatial Big Data, Machine Data, Spatio-Temporal Analytics, Point-Of-Care Diagnosis Devices

Back in 2001, Douglas Laney from Meta Group (an IT research company acquired by Gartner in 2005) wrote a research paper in which he stated that e-commerce had exploded data management along three dimensions: volumes, velocity, and variety (Laney, 2001).

Big data is defined as high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and processes optimization (Laney, 2012). The datasets in big data can be classified as structured or unstructured/semi-structured. Many experts believe that today more than 80% of all data is unstructured (Mayer and Cukier, 2014). Until recently unstructured data were mostly just stored without sophisticated means (and intention) for automatic analysis. Today the proliferation of big data tools and technologies has made it possible to manage and analyse any type of datasets to extract valuable insights automatically.

From another point of view, data component of big data can be broken down into two broad categories: human-generated and machine data. The human-generated data is the information this is generated when humans directly interact with an online system or record information, such as a social media (like emails, social networking application feeds, crowd-sourcing and/or collaborative content generation, and blogs) and click-stream data. At the other hand, machine data are generated directly by machines (without direct intervention of humans). For example, firewalls, load balancers, routers, switches, and computers (servers) that support human interactions with online systems, generate log files that describe the activities of user and the status of the system. In addition sensor-based data (like data generated automatically by smartphones for the monitoring of performance of mobile operating systems among the other things, data generated by jet engines for monitoring the health and performance of engine by engine manufacturer companies-and not the airline!) falls into the machine data category. Utilizing the mentioned machine data (by using efficient and automatic approaches) for research purposes largely ignored† until recently. However, machine data has vast and as yet

* Pouria.Amirian@ndm.ox.ac.uk

† Machine data has always been used for research purposes (like manual image processing of satellite images) and the automatic analysis of huge amount of machine generated data using supercomputers or crowd-source computing or volunteer computing has been done since late nineties. However the automatic analysis of machine data via bringing computing to data is new in big data world!

untapped potential as a source of highly valuable information, as they contain important insights about the system as well as users of the system.
Often data component of Big Data has a positional component as an important part of it in various forms, such as postal address, Internet Protocol (IP) address and geographical location. Given the above definition of big data, many researchers believe that geospatial data has been always big data!
As it defined by Pouria Amirian (Amirian et. al, 2014), "*geospatial Big Data"* term should be used if the positional components in Big Data extensively used in storage, retrieval, analysis, processing, visualization and knowledge discovery. In other words, geospatial Big Data systems need certain type of techniques, algorithms for efficient management, analytics and sharing based on best practices in management and analysis of geospatial data in the big data landscape.
As it mentioned by many researchers, management and analysis of geospatial data is complex and requires specific storage, processing, analysis and publication mechanisms (Taniar et. al, 2013; Amirian et. al, 2010; Mahboubi et. al, 2013; Basiri et. al, 2014; Yildizli et. al, 2011; Safar et. all, 2009; Basiri et. al, 2012). In fact management and analysis of geospatial data have been always revealed the limitations of information systems and computational frameworks. In a nutshell, unique characteristics of geospatial data such as high volume, various type of relationships between geospatial objects (e.g. distance, directional and topological relationships), need for long transactions, computationally intensive algorithms of processing and inclusion of time component, makes the management and analysis of geospatial Big Data even more complicated. Some researchers agreed that geospatial data may represent the biggest Big Data challenge of all (Davenport, 2014).
The focus of this paper is on finding useful insights from data generated by Point-Of-Care (POC) Diagnostics devices. In general the POC devices are sensors that provide results of various health-related tests. These devices are widely used in all types of healthcare settings and generate vast amount of machine data. Among various types of POC diagnostics devices, our research project has focused on data generated by advanced molecular POC diagnostic systems related to HIV and TB diseases.
Based on the mentioned definition of geospatial Big Data, at first glance data generated by the above molecular POC diagnostic devices seems not very relevant to geospatial Big Data. However by including location of the devices in the analysis which are usually locations of the healthcare settings (e.g. hospitals or labs) and by inclusion of other geographic layers, whole new set of useful questions can be asked and therefore many useful insights can be found. In order words if huge amount machine data (for example data about test results, resources used by devices, quality of the test, errors occurred during tests and location of devices) are available, using big data technologies it would be possible to answer the following important questions automatically:

- What is the demographic and geographic pattern of disease transmission? What is the spatio-temporal pattern of the disease prevalence? Is the rate of disease associated with geographic location, time or climate-change or special diet or ethnic group? How would be the spatio-temporal trend of prevalence of the disease (at population level) in future if the rate of spread is fixed as the current rate? (disease-related questions)
- What are the source of errors in data? Is there any statistically meaningful association between location and time with the type of errors for each device? Is the high rate of errors from specific Point-Of-Care devices meaningful? With the current rate of error how much the results of tests are reliable? Does the device maintained in expected manner? Does the device needs overhaul repair? Based on the current usage when is the best time for the next regular device inspection? (question related to the performance and health of the devices)
- What is the normal resource consumption for each location at specific time? Are the devices are used at their highest potential? With the current number of tests, which sites will need more/less resources (like cartridges) or even new POC devices than what is expected? What are the needed resources for the Point-Of-Care devices for next few months? (supply chain questions)

As presented by above sample questions, inclusion of geospatial data provides unprecedented

opportunities for using spatio-temporal data analysis. In addition using big data technologies enables to run the spatio-temporal data analysis in efficient manner. In This paper will describe the concept of geospatial Big Data from computational point of view and the explain the challenges, opportunities and findings of the research project currently being done at the Global Health Network, the Oxford University about using geospatial big data to find useful insights from machine data generated by POC devices.

References

Laney D (2001). 3D Data Management: Controlling Data Volume, Velocity and Variety.

Laney D (2012). The Importance of 'Big Data': A Definition.

Amirian P, Basiri A. and Winstanely, A (2014). *Evaluation of Data Management Systems for Geospatial Big Data*, Computational Science and Its Applications–ICCSA 2014, 678-690.

Mayer V and Cikier K (2014). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Eamon Dolan/Mariner Books, USA.

Taniar D and Rahayu W (2013). *A taxonomy for nearest neighbour queries in spatial databases*, Journal of Computer and System Sciences 79(7): 1017-1039.

Amirian P, Basiri A. and Alesheikh, A (2010). *Interoperable exchange and share of urban services data through geospatial services and XML database*, Complex, Intelligent and software intensive systems (CISIS), IEEE 62-68.

Mahboubi H, Bimonte S, Deffuant G, Chanet J and Pinet F (2013). *Semi-Automatic Design of Spatial Data Cubes from Simulation Model Results*. IJDWM 9(1): 70-95.

Basiri A, Amirian P and Winstanley A (2014). *The USE Of Quick Response (QR) Code In LandmarkBased Pedestrian Navigation*, International Journal of Navigation and observation.

Yildizli C, Pedersen T, Saygin Y, Savas, E, Levi, A (2011). *Distributed Privacy Preserving Clustering via Homomorphic Secret Sharing and Its Application to (Vertically) Partitioned SpatioTemporal Data*. IJDWM 7(1): 46-66.

Safar M, Ebrahimi D and Taniar D (2009). *Voronoi-based reverse nearest neighbor query processing on spatial networks*, Multimedia Systems, 15(5): 295-308.

Minelli M, Chambers M and Dhiraj A (2013). *Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses*, Wiley.

Basiri A, Amirian P, Winstanley A, Kuntzsch C, Sester M (2012). *Uncertainty handling in navigation services using rough and fuzzy set theory*, Proceedings of the Third ACM SIGSPATIAL International Workshop on Querying and Mining Uncertain Spatio-Temporal Data. ACM 38-41.

Davenport T H (2014). *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities*. Harvard Business Review Press.